Objective: To analyse how a customer's credit class is affected by their employment duration and type of job they are in.

Question 1: What is the relationship between years of employment and credit class?

Data Analysis Type	Descriptive, Diagnostic
Data Analysis Techniques	Descriptive, Factor (Cramer's V)
Independent Variable	Employment (Categorical)
Dependent Variable	Credit class (Categorical)
Data Visualization	Stacked bar chart, dodged bar chart

From the stacked bar chart in Figure 3.1.1, it can be seen that the credit class was mostly bad when employment was between one and four years but it was also mostly good for the same range of employment as demonstrated by the green squares. The second highest frequency for good classes was recorded for more than seven years of employment. The relationship could lean towards the higher the years of employment, the better the credit class. However, the credit class being mostly good when employment is "unemployed" and being mostly bad when it is between one and four years and between four and seven years (though the gap is not significant) goes against this logic. To get the actual frequencies, a dodge bar chart was also used.

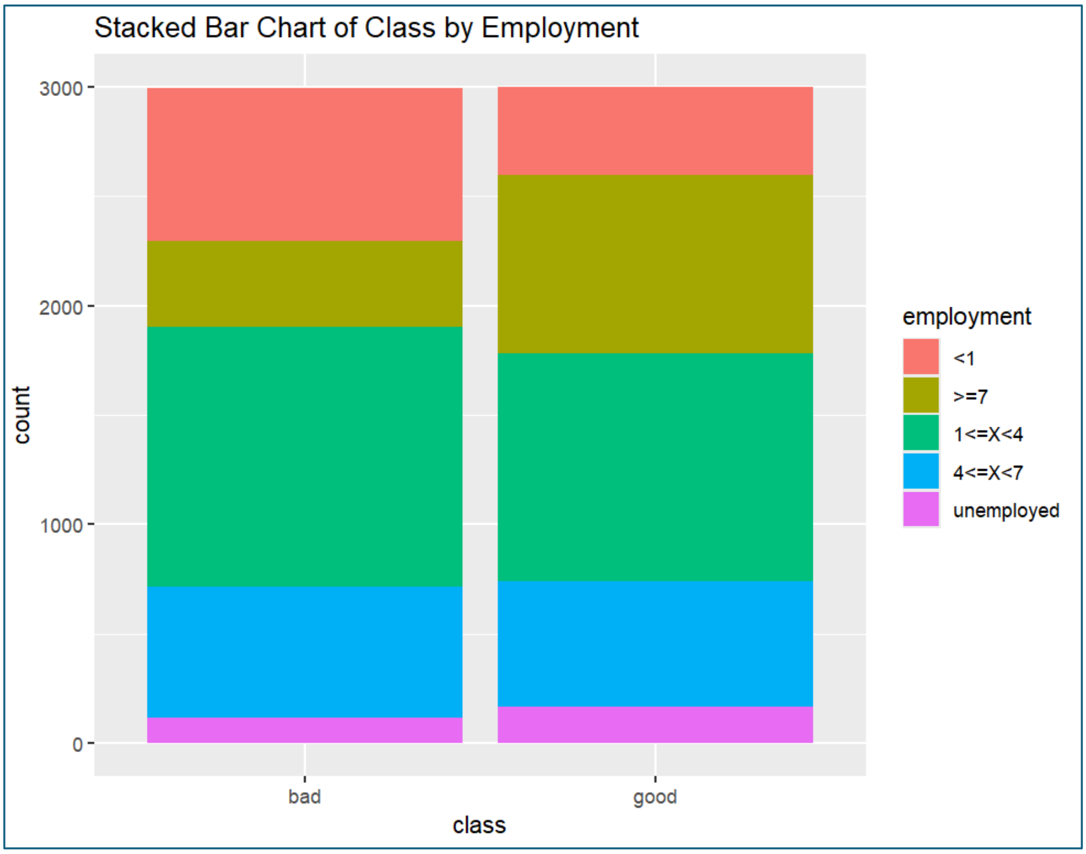

Figure 3.1.1: Stacked Bar Chart of Class by Employment

```
# stacked bar chart
ggplot(df, aes(x = class, fill = employment)) +
  geom_bar(position = "stack") +
  labs(title = "Stacked Bar Chart of Class by Employment ")
```

Figure 3.1.2: Code Snippet; Stacked Bar Chart of Class by Employment

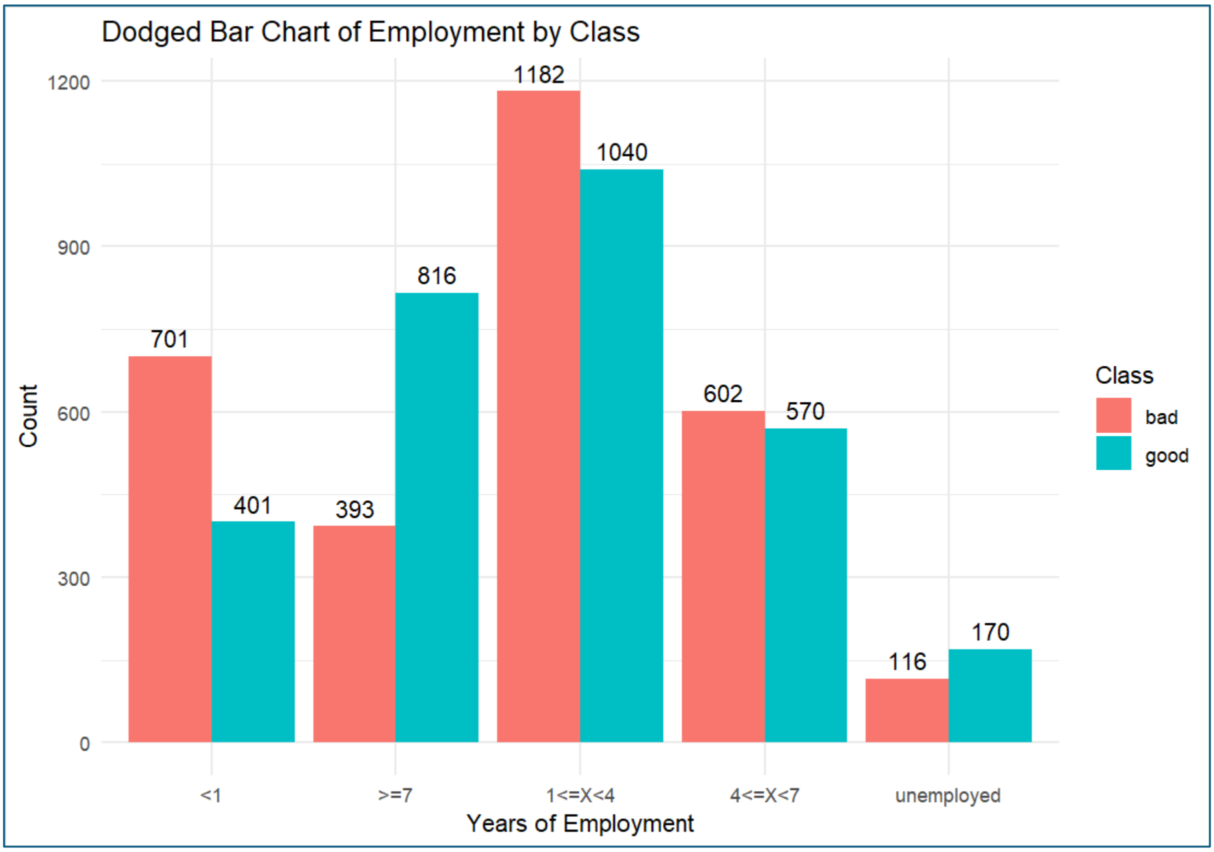

Figure 3.1.3: Dodged Bar Chart of Employment by Class

```
#Dodged bar chart of employment by class
#for all values of “employment”,
#how many corresponding values of “class” were good or bad.
ggplot(df, aes(x = employment, fill = class)) +
  geom_bar(position = "dodge") +  # make bars side-by-side
  geom_text(stat = 'count', aes(label = after_stat(count)), #use labels to view count
            position = position_dodge(width = 0.9), vjust = -0.5) +
  labs(title = "Dodged Bar Chart of Employment by Class",
       x = "Years of Employment",
       y = "Count",
       fill = "Class") +
  theme_minimal()
```

Figure 3.1.4: Code Snippet; Dodged bar chart of Employment by Class

To find the strength of the relationship between employment and credit class, Cramer’s V was used. Cramer’s V is a popular method of quantifying the relationship between two categorical variables. The Cramer’s V obtained was 0.204 denoting a slight to moderate relationship between employment and class. The null hypothesis is rejected as the Likelihood ratio and Pearson value are significantly large and the p value is 0. The years of employment can only slightly affect the credit class which explains why in certain cases, the credit class was bad despite several years of employment was recorded. It is therefore suggested that stakeholders look beyond employment stability when assessing the creditworthiness of clients.

```
> cramer_v
                    X^2 df P(> X^2)
Likelihood Ratio 254.10  4        0
Pearson          249.81  4        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.2
Cramer's V        : 0.204
>
```

Figure 3.1.5: Cramer's V results of Employment and Class

```
library(vcd) #import vcd library
#find the strength (instead of the nature)
#of the relationship between employment and credit class
#no relationship - cramer's v is 0
#perfect relationship - cramer's v is 1
#The higher the Pearson value, the more
#the null hypothesis (the variables have no relation) can be rejected

contingency_table <- table(df$employment, df$class) #create contingency table
cramer_v <- assocstats(contingency_table) #get cramer's v
cramer_v
```

Figure 3.1.6: Code Snippet; Calculating : Cramer's V of Employment and Class

Question 2: How is class affected when taking into account both employment and job?

Data Analysis Type	Descriptive, Diagnostic, Predictive, Prescriptive
Data Analysis Techniques	Descriptive, Factor (Cramer's V), Regression Analysis (Logistic Regression)
Independent Variable	Employment (Categorical), Job (Categorical)
Dependant Variable	Credit class (Categorical)
Data Visualization	Mosaic plot, stacked bar chart

1.1 Relationship between job and employment

It can be observed from the mosaic plot in Figure 3.1.7 that skilled jobs dominate the dataset. Irrespective of the years of employment, most workers in the dataset were in skilled jobs except for unemployed workers who were mostly in high quality and self-employed jobs. The latter observation may indicate errors in the dataset, some self-employed individuals may have been marked as "unemployed" under the "employment" variable. A summary of the mosaic plot findings can be found in Table 3.1.1 for better understanding.

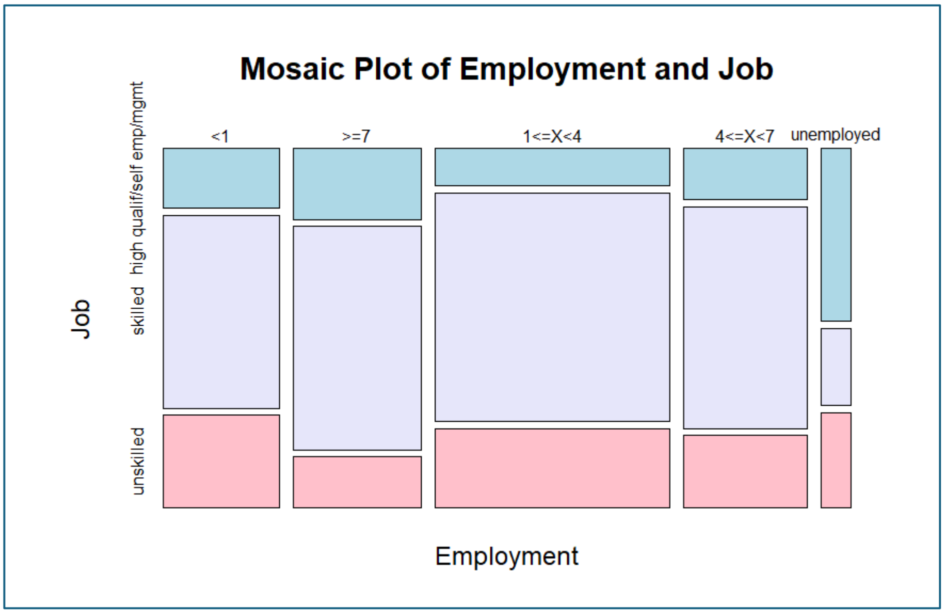

Figure 3.1.7: Mosaic Plot of Employment and Job

Years of employment	Most popular job type	Second most popular job type	Least popular job type
< 1	Skilled	Unskilled	High quality…
1<= X < 4	Skilled	Unskilled	High quality…
4 <= X < 7	Skilled	Unskilled	High quality
>= 7	Skilled	High quality…	Unskilled
unemployed	High quality…	Unskilled	Skilled

Table 3.1.1: Summary of Job Type by years of Employment

```
#MOSAIC PLOT of employment and job_______________________________________________
#As both are categorical variables, a mosaic plot is a good tool
#to plot the values of job and employment against each other.

color = c( "lightblue", "lavender", "pink")
#create contingency table first
contingency_table2 <- table(df$employment, df$job)
mosaicplot(contingency_table2, main = "Mosaic Plot of Employment and Job",
           xlab = "Employment", ylab = "Job", color = color)
```

Figure 3.1.8: Code Snippet; Plotting Mosaic Plot for Employment and Job

The Cramer's V obtained (0.18), however, denoted a rather weak relationship between the two variables. The years of employment of a worker may not accurately depict the type of job they are in.

```
                    X^2 df P(> X^2)
Likelihood Ratio 349.41  8        0
Pearson          386.18  8        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.246
Cramer's V        : 0.18
>
```

Figure 3.1.9: Cramer's V results of Employment and Job

```
contingency_table2 <- table(df$employment, df$job)
cramer_v <- assocstats(contingency_table2) #get cramer's v
cramer_v
```

Figure 3.1.10: Code Snippet; Calculating : Cramer's V of Employment and Job

1.2 Relationship between job and class

Here, a stacked bar chart was used at first to visualise how credit class varies with the type of job. Once again, skilled jobs dominate. Skilled jobs take up the biggest proportion out of all other jobs when it comes to both good and bad classes even though the proportion is bigger for good classes.

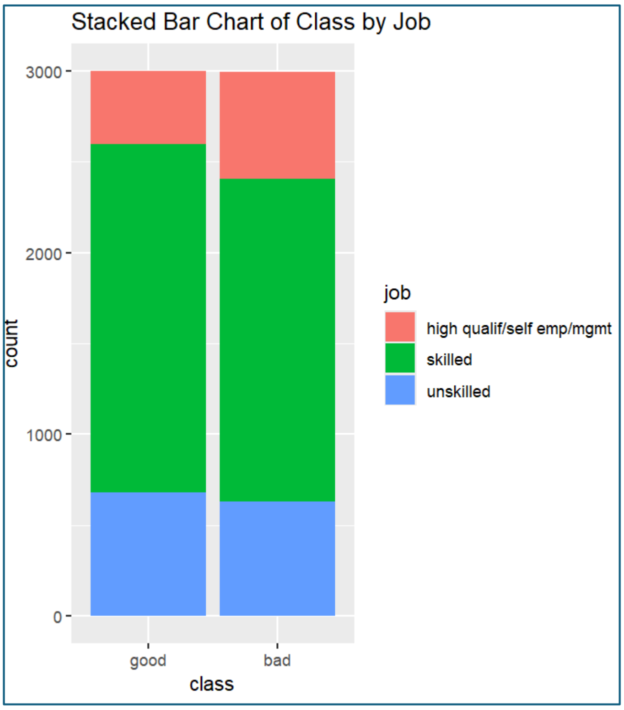

Figure 3.1.11: Stacked Bar Chart of Class by Job

```
#STACKED BAR CHART JOB AND CLASS
# stacked bar chart
ggplot(df, aes(x = class, fill = job)) +
  geom_bar(position = "stack") +
  labs(title = "Stacked Bar Chart of Class by Job")
```

Figure 3.1.12: Code Snippet; Stacked Bar Chart of Class by Job

Once again, the Cramer's V value was calculated to get the strength of the relation between job and class. A value of 0.085 was obtained denoting little to no relationship. Even the likelihood ratio and Pearson's values were relatively small as compared to earlier when the relationship between employment and class and employment and job were measured. This means that is less of deviation from the null hypothesis. The type of job a person has may not have a significant impact on their credit class as compared to the number of years for which they have been employed.

```
                    X^2 df   P(> X^2)
Likelihood Ratio 43.082  2 4.4134e-10
Pearson          42.866  2 4.9189e-10

Phi-Coefficient   : NA
Contingency Coeff.: 0.084
Cramer's V        : 0.085
> 
```

Figure 3.1.13: Cramer's V results of Job and Class

```
#get cramer's v for class and job
contingency_table3 <- table(df$class, df$job)
cramer_v <- assocstats(contingency_table3)
cramer_v
```

Figure 3.1.14: Code Snippet; Cramer's V results of Class and Job

2.3 <u>What would be the predicted class across different values of job and employment?</u>

To find out what the credit class would most likely be based on the type of job and employment duration, a logistic regression was used. The screenshot of the logistic regression model summary can be seen in Figure 3.1.16. As all p values were extremely small or close to 0 (denoted by ***) , all relations had very high statistical significance (employment and job are useful predictors) and the null hypothesis can be rejected. To view the coefficients (log odds) and odd ratios together, the broom and dplyr library was used. The output, which will be used for interpretation, can be seen in Figure 3.1.15. Log odds were exponentiated to get the odd ratios as the latter are easier to interpret.

1. **Job Type**

The positive coefficient means that both those in skilled and unskilled jobs are more likely to have a good credit class rather than a bad one compared to those in high quality, management jobs or those who are self-employed. For skilled jobs, the log odds ratio which is more than 1 (1.85) indicate that the odds of having a good credit class is 1.85 times as large as the odds for those with high quality jobs having a good credit class provided that employment is held constant. This can be explained by the frequency of skilled jobs in the dataset which was seen previously in the mosaic plot and stacked bar chart.

As for unskilled jobs, the results were quite surprising. The log odds ratio for unskilled jobs was 2.03 which indicates that the odds of having a good credit class is higher for unskilled workers compared to skilled workers and even higher than high quality workers. The difference between odds ratio is a small one (0.18) however this could suggest that other factors such as years of employment, savings status and more were more favourable for unskilled workers than skilled workers.

Generally, it can be concluded that those in low quality jobs (skilled and unskilled) were more likely to observe a good credit class than those in high quality jobs.

2. **Years of employment**

All coefficients were positive meaning that across all provided employment durations, workers were more likely to have a good credit class as compared to those having less than one year of employment. Workers with seven and more years of employment had the highest odds to fall in the good class category (3.93) followed by those who are unemployed (3.04) which was a surprising finding. However, the mosaic plot of employment and job clearly showed that majority of those registered as "unemployed" were in the "high quality/self employed/management" category. This could explain why the odds of having a good credit class are so high for the unemployed category. Another explanation could be that other factors such as savings status, credit history and more were promising despite the unemployed status. Moreover, as seen in our count plot, most unemployed workers had a good credit class (170 good credit class and 116 bad credit class).

The last two categories ranked from highest to lowest odds ratio are those having between four to less than seven years of employment (1.61) followed by those with one to less than four years of employment (1.55). The difference in odds ratio is not a significant one but the overall trend clearly suggests that longer employment tend to lead to a good credit class.

```
> formatted_findings
# A tibble: 7 × 6
  term                   estimate std.error statistic  p.value odds_ratio
  <chr>                     <dbl>     <dbl>     <dbl>    <dbl>      <dbl>
1 (Intercept)              -1.11     0.0952    -11.7  1.77e-31      0.329
2 employment>=7             1.37     0.0940     14.6  4.35e-48      3.93
3 employment1<=X<4          0.441    0.0802      5.50 3.77e- 8      1.55
4 employment4<=X<7          0.479    0.0908      5.27 1.35e- 7      1.61
5 employmentunemployed      1.11     0.149       7.48 7.67e-14      3.04
6 jobskilled                0.614    0.0817      7.52 5.43e-14      1.85
7 jobunskilled              0.706    0.0946      7.46 8.73e-14      2.03
>
```

Figure 3.1.15: Viewing Log Odds and Log Odds Ratio of the Logistic Regression model together

```
Call:
glm(formula = class ~ job + employment, family = binomial, data = training_set)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)          -1.06792    0.10111 -10.562  < 2e-16 ***
jobskilled            0.58610    0.08681   6.751 1.46e-11 ***
jobunskilled          0.68814    0.10013   6.872 6.31e-12 ***
employment>=7         1.34148    0.09994  13.423  < 2e-16 ***
employment1<=X<4      0.40840    0.08516   4.796 1.62e-06 ***
employment4<=X<7      0.46577    0.09607   4.848 1.25e-06 ***
employmentunemployed  1.08416    0.15769   6.875 6.18e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6644.5  on 4792  degrees of freedom
Residual deviance: 6388.0  on 4786  degrees of freedom
AIC: 6402

Number of Fisher Scoring iterations: 4
```

Figure 3.1.16: Summary of the Logistic Regression

```
#LOGISTIC REGRESSION___________________________________________________________

#make class ( the dependent variable) as a factor and relevel it so that the
#reference class is "bad"
df$class <- relevel(as.factor(df$class), ref = "bad")

set.seed(123)

#split dataset in 9:1 ratio
split = sample.split(df$class, SplitRatio = 0.9)

training_set = subset(df, split == T)

#test set is to use the trained model to predict on unseen data
test_set = subset(df, split == F)


#build logistic regression for model 1
classifier1 = glm(class ~ employment + job, training_set, family = binomial)
#model 1 summary
summary(classifier1)

#exponentiate the coefficients to get odds ratios
formatted_findings <- tidy(classifier1) %>%
  mutate(odds_ratio = exp(estimate))
formatted_findings
```

Figure 3.1.17: Code Snippet; Building a Logistic Regression model with Job and Employment as the independent variables and Class as dependent variable

Question 3: Are there any interactions between job and employment that could affect class? Does the impact of job on credit class depend on employment and vice versa?

Data Analysis Type	Diagnostic, Predictive, Prescriptive
Data Analysis Techniques	Factor, Regression Analysis (Logistic Regression)
Independent Variable	Employment (Categorical), Job (Categorical)
Dependant Variable	Credit class (Categorical)
Data Visualization	Partial Dependence Plot (PDP)
Machine Learning Model	Random Forest
Other	ANOVA test, Chi-Squared Distribution

The logistic regression model employed in 2.3 was fitted without any interaction term, hence the effects of job and employment on class are individually considered. However, it is possible for there to be interactions between the two variables. To assess whether there is a significant interaction between the two, an additional logistic regression was fitted this time with an interaction term.

Model 1: Without interaction

```
Call:
glm(formula = class ~ employment + job, family = binomial, data = training_set)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)          -1.11093    0.09518 -11.672  < 2e-16 ***
employment>=7         1.36894    0.09395  14.570  < 2e-16 ***
employment1<=X<4      0.44097    0.08016   5.501 3.77e-08 ***
employment4<=X<7      0.47855    0.09077   5.272 1.35e-07 ***
employmentunemployed  1.11051    0.14855   7.476 7.67e-14 ***
jobskilled            0.61426    0.08167   7.521 5.43e-14 ***
jobunskilled          0.70584    0.09463   7.459 8.73e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7474.9  on 5391  degrees of freedom
Residual deviance: 7173.7  on 5385  degrees of freedom
AIC: 7187.7

Number of Fisher Scoring iterations: 4
```

Figure 3.1.18: Summary of Logistic Regression Model 1: No interaction terms

```
> confusionMatrix(table(pred_class_test, test_set1$class))
Confusion Matrix and Statistics


pred_class_test bad good
           bad  193  157
           good 106  143

               Accuracy : 0.5609
                 95% CI : (0.5201, 0.6011)
    No Information Rate : 0.5008
    P-Value [Acc > NIR] : 0.001841

                  Kappa : 0.1221

 Mcnemar's Test P-Value : 0.002048

            Sensitivity : 0.6455
            Specificity : 0.4767
         Pos Pred Value : 0.5514
         Neg Pred Value : 0.5743
             Prevalence : 0.4992
         Detection Rate : 0.3222
   Detection Prevalence : 0.5843
      Balanced Accuracy : 0.5611

       'Positive' Class : bad
```

Figure 3.1.19: Confusion Matrix of Logistic Regression Model 1 following Prediction on a test set

Model 2: With interaction

```
Call:
glm(formula = class ~ employment * job, family = binomial, data = training_set)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)                        -1.5476     0.2011  -7.697 1.39e-14 ***
employment>=7                       1.4735     0.2428   6.068 1.30e-09 ***
employment1<=X<4                    0.9355     0.2459   3.805 0.000142 ***
employment4<=X<7                    1.2562     0.2571   4.885 1.03e-06 ***
employmentunemployed                1.9978     0.2693   7.418 1.19e-13 ***
jobskilled                          1.2099     0.2185   5.537 3.08e-08 ***
jobunskilled                        1.0010     0.2374   4.216 2.49e-05 ***
employment>=7:jobskilled           -0.2065     0.2707  -0.763 0.445651    
employment1<=X<4:jobskilled        -0.6936     0.2660  -2.607 0.009130 ** 
employment4<=X<7:jobskilled        -0.9536     0.2816  -3.387 0.000708 ***
employmentunemployed:jobskilled    -2.0211     0.3920  -5.156 2.52e-07 ***
employment>=7:jobunskilled          0.3670     0.3353   1.095 0.273637    
employment1<=X<4:jobunskilled      -0.2585     0.2923  -0.884 0.376521    
employment4<=X<7:jobunskilled      -0.8219     0.3175  -2.589 0.009637 ** 
employmentunemployed:jobunskilled  -0.8255     0.3940  -2.095 0.036140 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7474.9  on 5391  degrees of freedom
Residual deviance: 7126.7  on 5377  degrees of freedom
AIC: 7156.7

Number of Fisher Scoring iterations: 4
```

Figure 3.1.20: Summary of Logistic Regression Model 2:With interaction of Job and Employment

```
> confusionMatrix(table(pred_class_test2, test_set2$class))
Confusion Matrix and Statistics


pred_class_test2 bad good
            bad  225  185
            good  74  115

               Accuracy : 0.5676
                 95% CI : (0.5269, 0.6077)
    No Information Rate : 0.5008
    P-Value [Acc > NIR] : 0.0006139

                  Kappa : 0.1358

 Mcnemar's Test P-Value : 8.197e-12

            Sensitivity : 0.7525
            Specificity : 0.3833
         Pos Pred Value : 0.5488
         Neg Pred Value : 0.6085
             Prevalence : 0.4992
         Detection Rate : 0.3756
   Detection Prevalence : 0.6845
      Balanced Accuracy : 0.5679

       'Positive' Class : bad
```

Figure 3.1.21: Confusion Matrix of Logistic Regression Model 2 following Prediction on a test set

```
#PREDICTION
# Extract the desired columns
test_set1 <- test_set %>% select(employment, job, class)
test_set2 <- test_set %>% select(employment, job, class)


#predicting on test set using Model 1
pred_prob_test = predict(classifier1, type = "response", test_set1[,-3])
pred_class_test = ifelse(pred_prob_test > 0.5, "good", "bad")

confusionMatrix(table(pred_class_test, test_set1$class))

#add predicted class set to test set so comparison can be made
test_set1$pred_class <- pred_class_test
test_set1$pred_prob <- pred_prob_test
View(test_set1)
```

Figure 3.1.22: Code Snippet; Model 1 Predicting Class using test set

```
#interaction model - Model 2
#fit the model
classifier2 = glm(class ~ employment * job, training_set, family = binomial)
summary(classifier2)
#predict on test set using Model 2
pred_prob_test2 = predict(classifier2, type = "response", test_set2[,-3])

pred_class_test2 = ifelse(pred_prob_test2 > 0.5, "good", "bad")


confusionMatrix(table(pred_class_test2, test_set2$class))

#add predicted class set to test set so comparison can be made
test_set2$pred_class <- pred_class_test
test_set2$pred_prob <- pred_prob_test
View(test_set2)
```

Figure 3.1.23: Code Snippet; Building Model 2 with Interaction between Job and Employment and Predicting Class using test set

From Model 2's summary, only a few interactions had statistical significance, and the effect was negative for all except for the more than seven and unskilled interaction but the latter was not statistically significant. As for predictions, Model 2 (0.5676) has a slightly better accuracy than Model 1 (0.5609) indicating slight interaction between employment and job which when considered could improve the prediction of credit class.

Comparing sensitivity and specificity values, Model 2 was slightly better at correctly identifying bad credit classes but struggled when it came to good credit class cases. This could suggest that the interaction between job and employment was stronger in cases where the actual credit class was bad. To debate this hypothesis, a random forest with feature importance was conducted on the same training test. Indeed, the SHAP (SHapley Additive exPlanations) values for employment and job, seen in Figure 3.1.24, reveal that the combination of the two contribute more towards the prediction of bad classes than good classes. This suggests that employment and job, together, are strong and influential factors when predicting bad classes.

```
> importance_matrix
                              bad     good MeanDecreaseAccuracy MeanDecreaseGini
checking_status         103.72743 97.14905            108.23051        491.93199
duration                 62.57330 66.84267             67.71540        267.29461
credit_history           53.13577 51.97167             58.13996         86.53564
purpose                  54.31033 57.52748             57.61086        206.67068
credit_amount            79.62702 73.45944             86.12339        288.44620
savings_status           48.30047 47.86385             51.08124         91.96587
employment               57.17038 58.10636             63.88113        106.48934
installment_commitment   60.09600 58.62458             66.14436         94.56831
personal_status          47.70551 44.23074             47.76694        102.78678
other_parties            33.50157 35.27123             39.42873         33.45089
residence_since          58.92503 61.26385             63.91635        142.85858
property_magnitude       51.58570 51.88882             54.62204        140.53456
age                      60.85659 65.20403             66.83354        212.35894
other_payment_plans      33.36084 37.79497             37.25591         54.64959
housing                  40.90730 34.72508             41.61228         59.19961
existing_credits         41.38767 36.50747             40.82356         95.24831
job                      50.14660 49.16734             57.21152         59.35840
num_dependents           36.22233 35.03784             39.64031         48.68837
own_telephone            34.99988 34.12938             36.04557         63.18444
foreign_worker           27.94638 26.22495             29.42285         28.72716
```

Figure 3.1.24: Importance Matrix of Random Forest Model

```
#Random Forest - Feature Importance

#perform hot deck to replace missing values based on class value
training_set = hotdeck(training_set, domain_var = "class", imp_var = FALSE)

rf_model <- randomForest(x = training_set[, -21],
                         y = training_set$class,
                         ntree = 500,
                         importance = TRUE)

importance_matrix <- importance(rf_model)
importance_matrix
```

Figure 3.1.25: Code Snippet; Building a Random Forest Model with Feature Importance on Training set with all Variables as Predictors

<u>3.1 Comparison of the two models: how far is the interaction model a better predictor of class?</u>

To further compare the two logistic regression models, an ANOVA test was performed. The ANOVA results indicate that Model 2 is a better fit to the data due to its lower residual deviance which is 47.084 lower than Model 1. Regardless, the residual deviance is very high for both models. This can be explained by the fact that the regression models only accounts for two variables only while the credit class is affected by all the variables in the dataset with some having much bigger influence.

Furthermore, to determine if the improvement in the goodness of fit in Model 2 is statistically significant, the deviance was compared to a chi-squared distribution. The p-value obtained can be seen in Figure 3.1.27. As the p-value is less than 0.05, the improvement in fit provided by the interaction term (employment * job) in Model 2 is considered statistically significant.

```
> anova(classifier1, classifier2)
Analysis of Deviance Table

Model 1: class ~ employment + job
Model 2: class ~ employment * job
  Resid. Df Resid. Dev Df Deviance
1      5385     7173.7
2      5377     7126.7  8   47.084
> 
```

Figure 3.1.26: Results of ANOVA test on Model 1 and Model 2

```
> p_value <- pchisq(47.084, df = 8, lower.tail = FALSE)
> p_value
[1] 1.477863e-07
```

Figure 3.1.27: p-value of Chi-Squared Distribution

```
#anova of model 1 and model 2
anova(classifier1, classifier2)

#chi squared distribution
#47.084 = difference in deviance of model 1 and 2
#degrees of freedom = 8
p_value <- pchisq(47.084, df = 8, lower.tail = FALSE)
p_value
```

Figure 3.1.28: Code Snippet; Performing ANOVA test and Getting p-value of the Chi-Squared Distribution

3.2 Visualising the interaction

Finally, the interaction between employment and job was plotted against the predicted probabilities of class using the "effects" package and can be seen in Figure 3.1.29. As the gradients across the lines differ, it can be concluded that there is an interaction between employment and job.

The credit class when the job type is high quality seems to deviate a lot from other job types except for when the employment duration is between four to seven years, at this value of employment, the credit class across all job types seem very close to each other. When years of employment surpass seven however, skilled and unskilled jobs are once again at a bigger advantage and have their credit class improve significantly while that for high quality jobs only improve slightly.

The credit class of high-quality jobs seem to improve the most when employment is marked as "unemployed" while for skilled and unskilled jobs, the credit class is at their second time highest. This was a constant observation throughout the analysis. As seen in the bar chart in Figure 3.1.3 or log ratios of previous logistic regression models, despite being unemployed, the credit class was or was predicted to be good respectively. When it comes to high quality jobs, the reason, as mentioned before, could be because the high-quality jobs involve self-employed jobs. Overall, reasons could well include the data cleaning method used or errors when compiling the dataset, and the fact that a higher proportion of those marked as "unemployed" observed a good credit class compared to bad in the dataset.

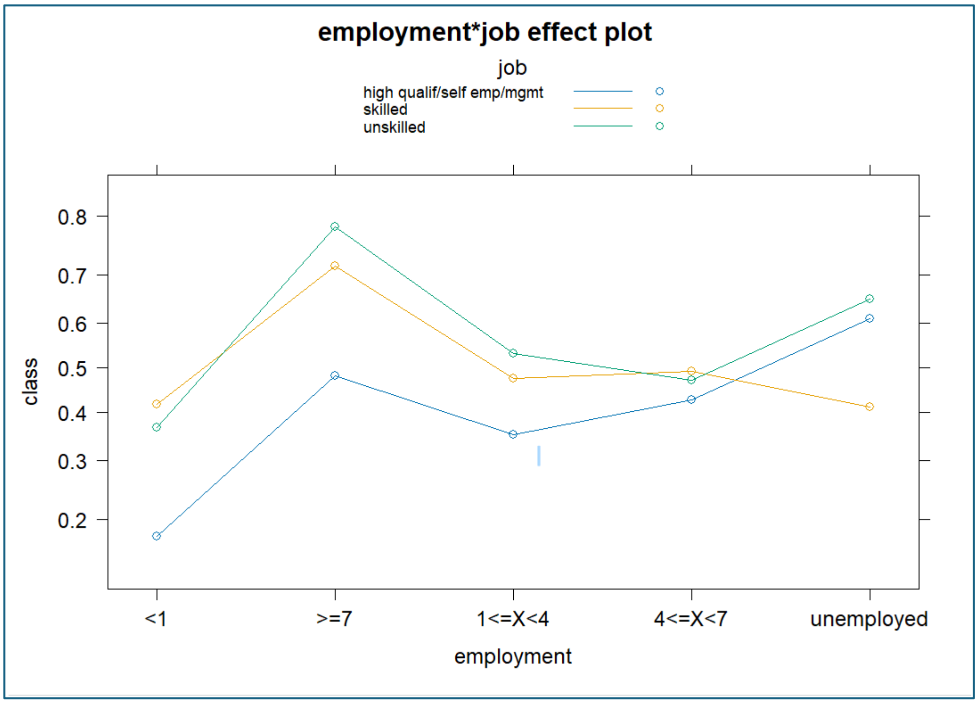

Figure 3.1.29: Plot of Interaction Effects in Model 2

```
plot(allEffects(classifier2), multiline = TRUE, ci.style = "bands")
```

Figure 3.1.30: Code Snippet; Plotting Interaction Effects

The PDP (Partial Dependence Plot) for job and employment was then plotted to find the marginal effect each variable has on the predicted outcome of Model 2. A PDP isolates all other interactions and only considers the effect of the included feature on the predicted outcome. We can thus find the separate effect of job and employment on the predicted class of Model 2 and then compare the observations with a two-feature PDP to precisely view how the predicted class changes when the effects of job and employment are considered together.

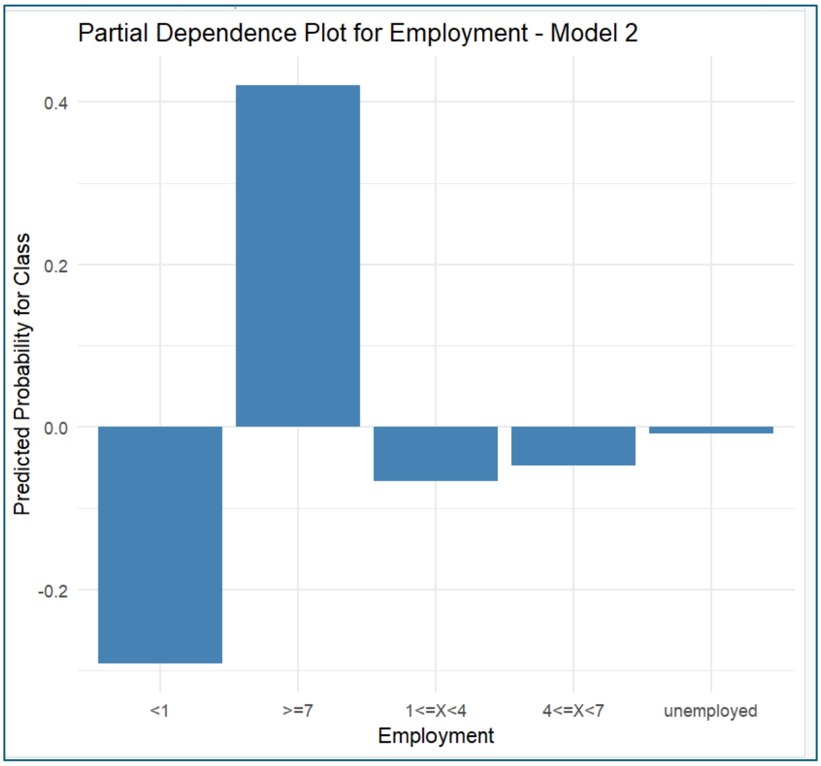

Figure 3.1.31: PDP of Employment

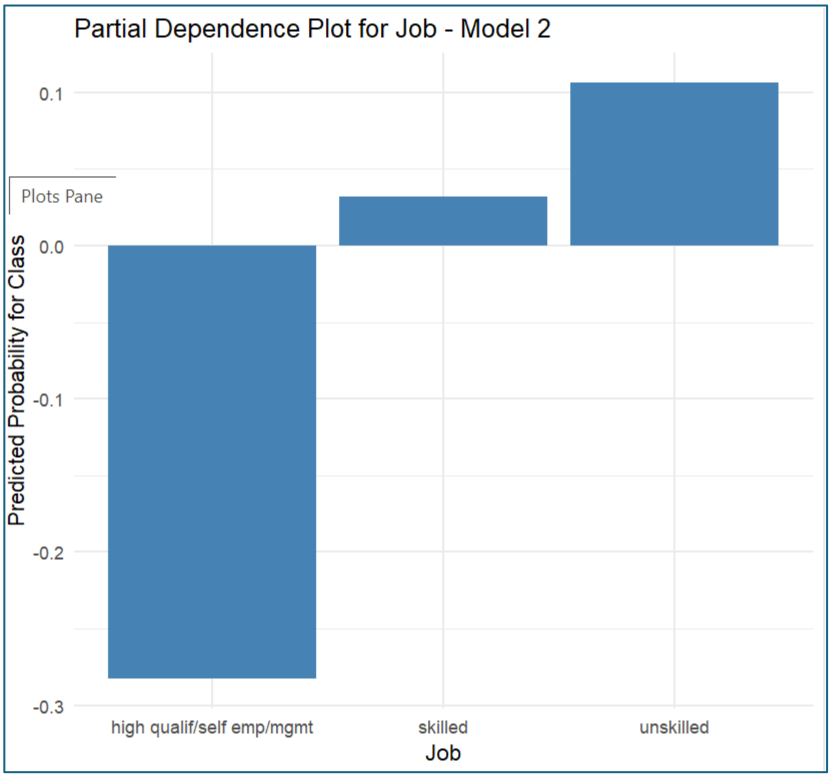

Figure 3.1.32: PDP of Job

Table 3.1.2: Summary of the PDP of Employment

Employment Duration	**Average Predicted Probability for Class (x = Predicted Probability)**	**Effect on Predicted Class***
< 1	-0.2 < x < -0.3	Worsens
>= 7	0.4 < x < 0.5	Improves
1 <= X < 4	0 < x < -0.1	Worsens
4 <= X < 7	0 < x < -0.1	Worsens
unemployed	0 < x < -0.1	Worsens

Table 3.1.3: Summary of the PDP of Job

Job Type	Average Predicted Probability for Class (x = Predicted Probability)	Effect on Predicted Class*
High Quality	-0.25 < x < -0.35	Worsens
Skilled	0 < x < 0.05	Improves
Unskilled	0.1 < x < 0.15	Improves

*Improves – Increase in likelihood of a good class being predicted, Worsens – Decrease in likelihood of good class being predicted

The results of the logistic regression, Model 1, inferred that workers in higher employment durations were more likely to have a good credit class as compared to those having less than one year of employment. The PDP of employment shows exactly how the predicted class probability would change and confirms the observation of Model 1. Even though the predicted probability improves, the average probability is still negative for all employment types except for ">=7".

For job, however, the PDP clearly shows a significant and positive improvement in predicted probability when moving from a high-quality job.

The two-feature PDP further details the interaction between job and employment. For instance, when employment is >= 7, the isolated effect on class leans towards a good class however when job type is high quality, it worsens the probability and when it is unskilled, it improves the class and increases the probability to above 0.5. Employment duration of four to seven years may be seen as stagnant years as the job type had little to no effect on the average predicted class. The effect of all interactions on the predicted class was summarised in Table 3.1.4.

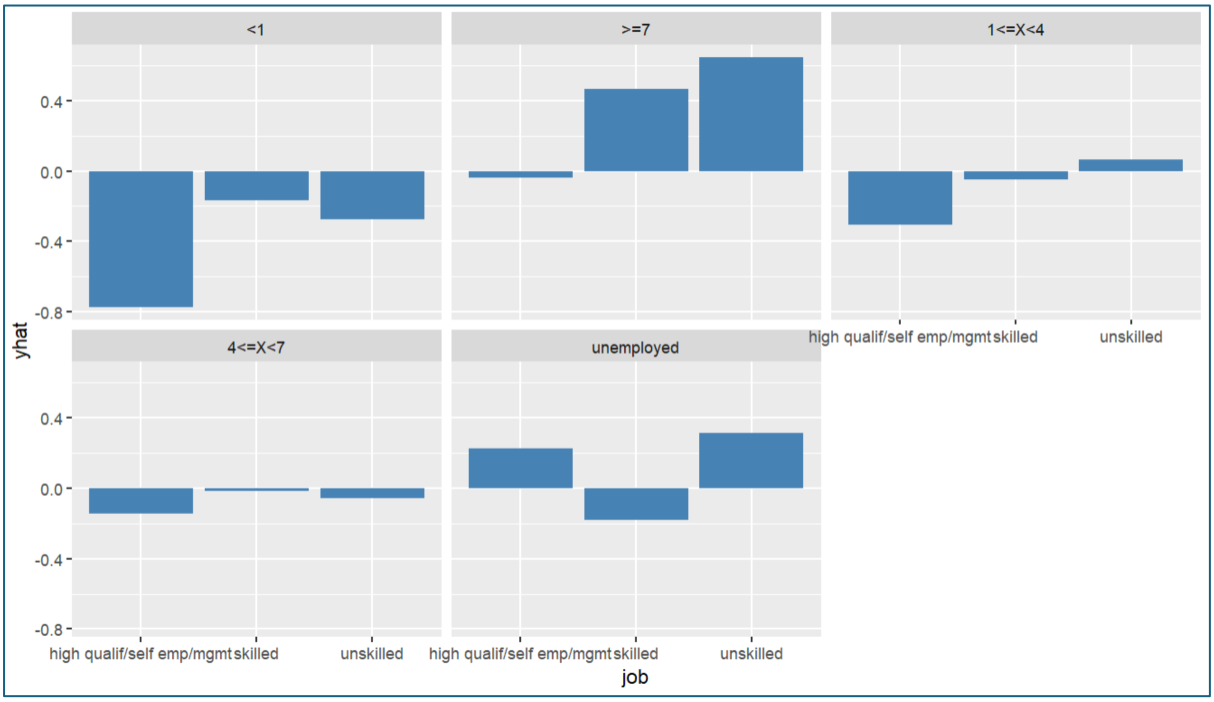

Figure 3.1.33: Two-Feature PDP of Job and Employment

Table 3.1.4: Summary of Two Feature PDP

Employment Duration	**Job Type**	**Combined effect on Predicted Class** (in comparison to isolated effect of employment duration)	**Average Predicted Probability for Class (x = Predicted Probability)**
< 1	High Quality	Worsens	-0.6 < x< -0.8
	Skilled	Slight Improvement	0 < x < 0.2
	Unskilled	No significant change	-0.2 < x < -0.4
>= 7	High Quality	Worsens	0 < x < -0.2
	Skilled	No significant change	0.4 < x< 0.6
	Unskilled	Improves	0.6 < x < 0.8
1 <= X < 4	High Quality	Slight worsening	-0.2 < x < -0.4
	Skilled	No significant change	0 < x < -0.2
	Unskilled	Slight improvement	0 < x < 0.2
4 <= X < 7	High Quality	No significant change	0 < x < -0.2
	Skilled	No significant change	0 < x < -0.1
	Unskilled	No significant change	0 < x < -0.2
unemployed	High Quality	Improves	0.2 < x < 0.3

	Skilled	No significant change	0 < x < -0.2
	Unskilled	Improves	0.2 < x < 0.4

As such, the most appropriate conclusion is that the effect of employment on predicted credit class depends on job type and vice versa with the effects being stronger when job type is either high quality or unskilled. Stakeholders are recommended to consider the employment duration and job type of an individual together for a better prediction of their creditworthiness. The current economic situation should also be considered as the prejudice about lower quality jobs may lead to incorrect predictions as the analysis made clearly shows that unskilled or skilled workers can have a better credit class over those in high quality or self-employed jobs in some cases.

```
#Partial Dependent Plot (PDP)
#The partial function tells us for each given value of job and employment
#what the average marginal effect on the prediction is by
#replacing the job type/ employment category of all data instances
#with each possible value and averaging  the predictions

#a flat PDP indicates that the feature
#is not important, and the more the PDP varies, the more
#important the feature is.


#show how each employment category/job category influences the
#model's predicted outcome while averaging over the effects
#of other features.


#turn employment and job to factor so pdp can be plotted
training_set$employment <- factor(training_set$employment)
training_set$job <- factor(training_set$job)
table(training_set$class)

classifier2 = glm(class ~ employment * job, training_set, family = binomial)
```

Figure 3.1.34: Code Snippet; Plotting PDPs of Employment and Job(1)

```r
#pdp of employment
pdp_emp <- partial(classifier2, pred.var = "employment", grid.resolution = 50)

ggplot(pdp_emp, aes(x = employment, y = yhat)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Partial Dependence Plot for Employment - Model 2",
       x = "Employment",
       y = "Predicted Probability for Class") +
  theme_minimal()

#pdp of job
pdp_job <- partial(classifier2, pred.var = "job", grid.resolution = 50)

ggplot(pdp_job, aes(x = job, y = yhat)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Partial Dependence Plot for Job - Model 2",
       x = "Job",
       y = "Predicted Probability for Class") +
  theme_minimal()
```

Figure 3.1.35: Code Snippet; Plotting PDPs of Employment and Job (2)

```r
#two-variable PDP shows the dependence of the
#class on joint values of job and employment
#can help identify possible interactions

pdp_result2 <- partial(classifier2, pred.var = c("job", "employment"))

ggplot(pdp_result2, aes(x = job, y = yhat)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  facet_wrap(~ pdp_result2$employment)
labs(title = "PDP of Logistic Regression Model 2",
     x = "Employment",
     y = "Predicted Probability for Class") +
  theme_minimal()
```

Figure 3.1.36: Code Snippet; Plotting Two-Feature PDP of Employment and Job