Job satisfaction in the industrial sector: A survey through questionnaires

Nikolaos Gkiouzelis
Informatics Department
Aristotle University of Thessaloniki
Thessaloniki, Greece
ngkiouzel@csd.auth.gr

Efstratios Grigoroudis
Informatics Department
Aristotle University of Thessaloniki
Thessaloniki, Greece
egrigorou@csd.auth.gr

Konstantinos Avgitidis
Informatics Department
Aristotle University of Thessaloniki
Thessaloniki, Greece
kavgitidi@csd.auth.gr

ABSTRACT

Statistical data analysis is a procedure of performing various statistical operations to identify trends and patterns that can be used for a plethora of options like predicting an outcome. The research topic that we focused, is predicting the job satisfaction of the employees in the industrial sector through a set of common independent variables and using a different independent variable for each of our models. The analysis contains extended results to answer a high portion of research questions on relationships between the dataset's variables. Moreover, we developed simple and more complex models presenting the ones with greater statistical significance.

KEYWORDS

Statistical data analysis, job satisfaction, relationship detection, descriptive analysis, logistic regression

1 Introduction

Job satisfaction is a vital element of the employee's performance nowadays. It is affected by several elements namely income, working hours per week, colleague's teamwork, and morale and many more. In the current study an extensive statistical analysis about this manner was conducted. The authors were given 747 answered questionnaires about job satisfaction. With the utilization of the R programming language various data preprocessing techniques, and descriptive analysis of the data were implemented. Furthermore, the correlations between the data variables were explored and the survey was concluded with some models that tried to captivate the trends of the data and to produce decent results.

The rest of the paper is structured as follows:

- In section 2 we present the data and the data preprocessing frameworks analytically.
- In section 3 we display the descriptive analysis of the data to obtain a more intuitive view of the current dataset.
- In section 4 we explain the correlations between the dependent and independent variables and hint how they may affect the performance of the models.
- In section 5 the best performing models are presented and analyzed.
- Finally, in section 6 we conclude the paper presenting several key conclusions that were obtained.

1.1 Key research questions

In this section we will briefly address the key research questions that are explored through this statistical data analysis. More specifically in this framework we aim to shed some light in the following research questions:

- Do the parameters age, education, sex, weekly working time, income, and importance of having a fulfilling job of the questioned respondents influence their job satisfaction?
- Do the parameters age, education, sex, weekly working time, job satisfaction and importance of having a fulfilling job of the questioned respondents influence their income when it is divided in 4 categories?
- Do the parameters age, education, sex, weekly working time, job satisfaction and importance of having a fulfilling job of the questioned respondents influence their income when it is divided in 18 categories?

2 Data

The given dataset contains 19 variables, from which some of them are categorical and some are numeric in terms of type. The nature of the variables allows us to assume that several variables could potentially play the role of the dependent variable. More specifically we considered satjob2, income4 and rincom91 as dependent variables and we implemented a modelling framework for every one of them considering at the same time all the rest of the variables as independent. A detailed description of the variables is provided in Table 1.

Variable / Variable type	Description	Levels – Values
age / numeric	Age of the respondent	98 = "DK" 99 = "NA"
educ / numeric	Highest year of school completed	97 = "NAP" 98 = "DK" 99 = "NA"
sex / categorical	Respondent's sex	1 = "Male" 2 = "Female"
degree / categorical	Respondent's highest degree	0 = "Less than HS" 1 = "High school" 2 = "Junior college" 3 = "Bachelor" 4 = "Graduate" 7 = "NAP" 8 = "DK" 9 = "NA"
satjob / categorical	Job satisfaction	0 = "NAP" 1 = "Very satisfied" 2 = "Mod satisfied" 3 = "A little dissatisfied" 4 = "Very dissatisfied" 8 = "DK" 9 = "NA"
satjob2 / categorical	Job satisfaction	1 = "Very satisfied" 2 = "Not very satisfied"
income4 / ordinal	Respondent's income	1 = "24.999 or less" 2 = "25.000 to 39.999" 3 = "40.000 to 59.999" 4 = "60.000 or more"
rincom91 / ordinal	Respondent's income	0 = "NAP" 1 = "Lower than $1000" 2 = "$1.000-2.999" 3 = "$3.000-3.999" 4 = "$4.000-4.999" 5 = "$5.000-5.999" 6 = "$6.000-6.999" 7 = "$7.000-7.999" 8 = "$8.000-9.999" 9 = "$10.000-12.499" 10 = "$12.500-14.999" 11 = "$15.000-17.499" 12 = "$17.500-19.999" 13 = "$20.000-22.499" 14 = "$22.500-24.999" 15 = "$25.000-29.999" 16 = "$30.000-34.999" 17 = "$35.000-39.999" 18 = "$40.000-49.999" 19 = "$50.000 - 59.999" 20 = "$60.000-74.999" 21 = "$75.000+" 22 = "REFUSED" 98 = "DK" 99 = "NA"
hrs1 / numeric	Number of hours worked last week	-1 = "NAP" 98 = "DK" 99 = "NA"
wrkstat / categorical	Labor force status	0 = "NAP" 1 = "Working fulltime" 2 = "Working part-time" 3 = "Temp not working" 4 = "Unemployed, laid off" 5 = "Retired" 6 = "School" 7 = "Keeping house" 8 = "Other" 9 = "NA"
jobinc / categorical	Importance of high income	0 = "NAP" 1 = "Most important" 2 = "Second" 3 = "Third" 4 = "Fourth" 5 = "Fifth" 8 = "DK" 9 = "NA"
impjob / categorical	Importance to R of having a fulfilling job	0 = "NAP" 1 = "One of the most important" 2 = "Very important" 3 = "Somewhat important" 4 = "Not too important" 5 = "Not at all important" 8 = "DK" 9 = "NA"
bothft / categorical	both spouses work full time	0 = "No" 1 = "Yes"
husbft / categorical	Husband employed full time	0 = "No" 1 = "Yes"
husbhr / numeric	hrs worked last week by husband	None
wifeft / categorical	Wife employed full time	0 = "No" 1 = "Yes"
wifehr / numeric	hrs worked last week by wife	None
id		None
agecat4 / categorical	4 categories of age	1 = "18-29" 2 = "30-39" 3 = "40-49" 4 = "50+"

Table 1: Data description

Additionally, the dataset had several missing values. Most of them existed only in specific variables namely husbft, husbhr, wifeft, wifehr which led us to different preprocessing techniques in order to extract the maximum possible information from this dataset.

2.1 Data preprocessing frameworks

As mentioned previously, because of the missing values of the dataset we concluded to split our work in two main data preprocessing techniques as follows:

- In framework 1 we contain all the variables, and we omit all the missing values concluding to a final dataset of 200 answered questionnaires which accounts for the 26.7% of the initial dataset
- In framework 2 we omit the 4 variables with the most missing values namely husbft, husbhr, wifeft, wifehr and then after we discard all the missing values resulting in a dataset of 460 answered questionnaires which accounts for the 61.5% of the initial dataset. Furthermore, we aggregate some categories of the variable's degree and rincom91 to optimize their information extraction.

The key difference between framework 1 and 2 is that framework 1 utilizes all the available variable information sacrificing some questionnaires whereas framework 2 has severely more data in its disposal but loses the trends that the 4 omitted variables might hide.

3 Descriptive analysis

In this section we will describe in detail the numeric and categorical variables of the aforementioned dataset, implementing several figures and analyzing them extracting many key conclusions about our variables.

Age and hrs1 variables have an expanded distribution of values resulting in several outliers in the boxplots presented in Figure 1 whereas education's values are concentrated around it's mean. This might indicate that the values of age and hrs1 variables do not follow a normal distribution, but we will comment this fact later in Figure 5 in more detail.

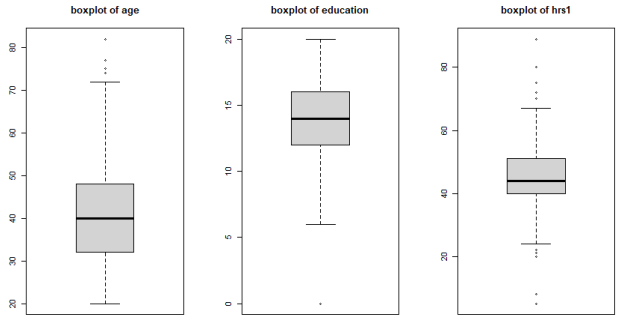

Figure 1: Boxplots of the numeric variables age, education and hrs1

From the histograms in Figure 2 we can pinpoint that the vast majority of values of the age variable are in the interval [25,55], the vast majority of the education variable values are in the interval [10,17.5] and the vast majority of the hrs1 variable values are in the interval [30,60].

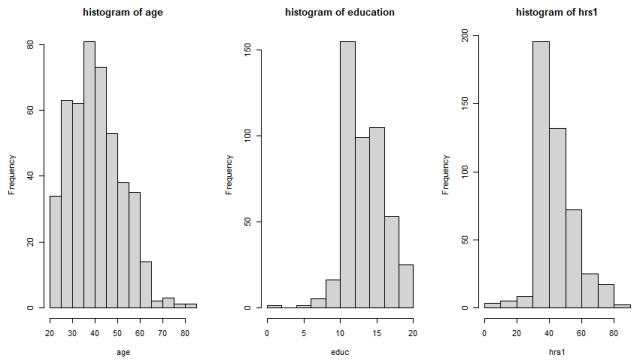

Figure 2: Histograms of the numeric variables age, education and hrs1

The variable degree in Figure 3 and rincom91 in Figure 4 are aggregated in two categories to utilize them better in the correlation and modeling upcoming sections. Some key conclusions that can be drawn from Figures 3,4 are the following:

- Most of the questionnaires were answered by men
- Most of the people had less than graduate degree
- Most of the people were at least moderate satisfied with their current jobs but when they were asked to categorize this moderate answer even further in a yes/no query they admitted that they tend not to be very satisfied as we can conclude from the satjob and satjob2 bar plots
- Most of the husbands or wives are mostly not working fulltime as we can conclude from the bothft bar plot
- Most of the people consider their job at least third in importance in their life or at least very important
- Most of the people are paid with less than 60.000$

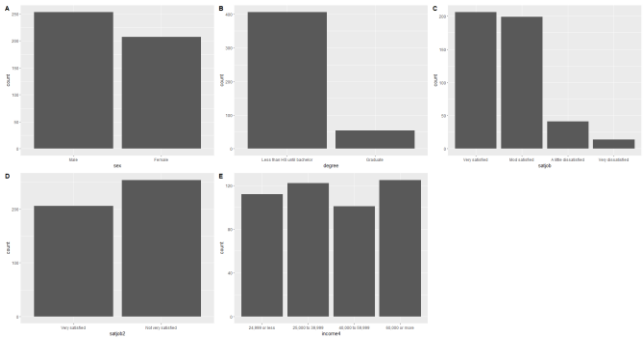

Figure 3: Bar plots of the variables sex, degree, satjob, satjob2, income4, jobinc, impjob, and bothft

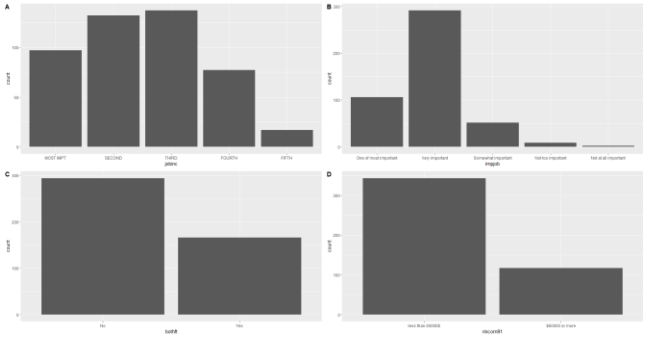

Figure 4: Bar plot of the variables rincom91, jobinc, impjob, and bothft

As we expected from Figure 1 results Figure 5 also indicates that both age, education and hrs1 are not following a normal distribution. Transforming the values of age and hrs1 with the logarithmic transformation, Figure 6, helps the value to follow the line better but still we can't say that either of their values follow a normal distribution. Log(educ) could not stand due to zero values in the education variable and other transformation like exp(educ) or educ^2 did not help at all transforming its values into a form towards normal distribution.

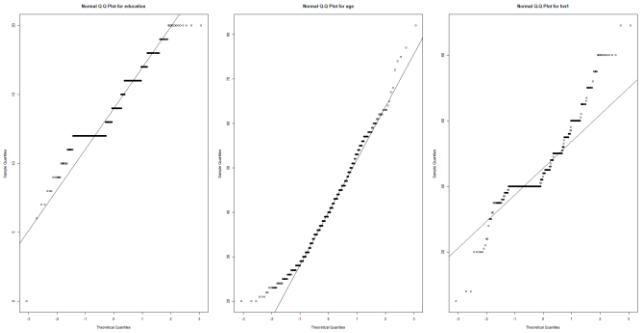

Figure 5: Q-Q plots of the numeric variables age, education and hrs1

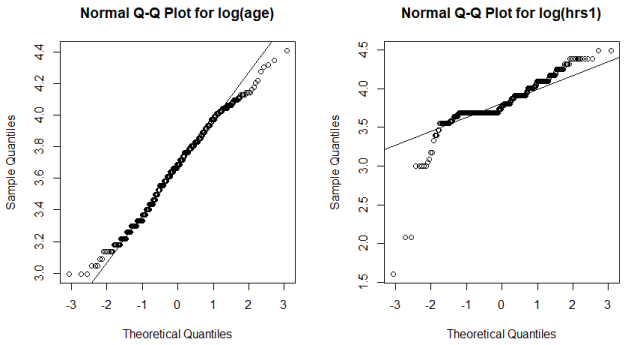

Figure 6: Q-Q plots of the numeric transformed variables log(age) and log(hrs1)

4 Correlations

4.1 Correlations between satjob2 (dependent) and the rest of variables (independent)

In this section we will examine the relationships between the dependent variable satjob2 and the rest of the variables which are considered as independent in this case study.
We observe both from Tables 2 and 3 and Figure 7 that the correlation between satjob2 and the numeric variables age, educ and hrs1 are weak. This fact indicates that these variables will not help the modeling procedure of the dependent satjob2 variable at a great effect.

Biserial correlation values	Age	Educ	Hrs1
Satjob2	0.09	0.08	0.06

Table 2: Biserial correlation between the dependent variable satjob2 and the independent numeric variables age, educ and hrs1

Wilcoxon p-values	Age	Educ	hrs1
Satjob2	0.059	0.068	0.18

Table 3: The resulting p-values of the conducted Wilcoxon test between the dependent variable satjob2 and the independent numeric variables age, educ and hrs1

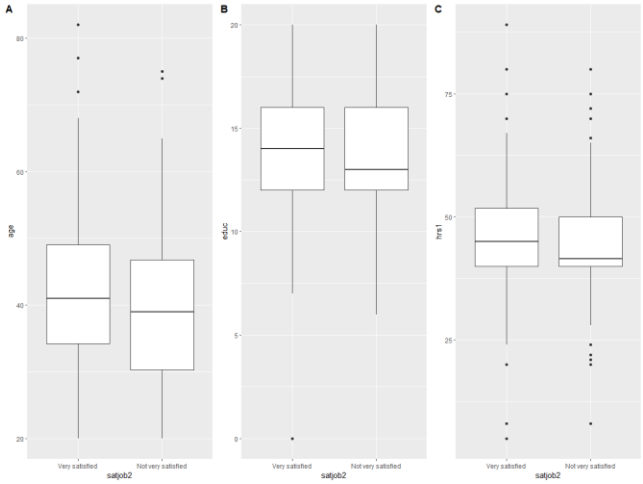

Figure 7: Boxplots of the correlation between the dependent variable satjob2 and the independent numeric variables age, educ and hrs1

In Table 4 we see that several categorical variables seem to have a statistically significant correlation between the independent variable satjob2 namely degree, satjob, income4, rincom91 and jobinc. Nevertheless, satjob has the same intuition as satjob2 and it can't be considered to the modeling framework of the dependent variable satjob2 because it will lead to overfitting results, and it would be like trying to predict modeling satjob2 by itself. Finally, wrkstat has only one category namely working fulltime and so we can't conduct any statistical tests and it would have no meaning to contain wrkstat to the model.

Chisq test p-values / Fisher test p-values	Satjob2
Sex	0.9/0.93
Degree	**0.047**/0.057
Satjob	**$<2.2*10^{-16}$ / $<2.2*10^{-16}$**
Income4	**0.002/0.002**
Rincom91	**0.0054/0.004**
Jobinc	**0.013/0.011**
Impjob	**0.036/0.026**
Bothft	0.95/1
Agecat4	0.16/0.15

Table 4: The resulting p-values of the conducted chi squared and fisher test between the dependent variable satjob2 and the independent variables sex, degree, satjob, income4, rincom91, wrkstat, jobinc, impjob, bothft and agecat4

Table 5 presents analytically the total amount of values in the 2 categories of the bothft variable in terms of the categories of satjob2. The non-correlation of these 2 variables is firmly proven pinpointing the fact both categories of the bothft variable have the same distribution of their values in the two categories of the satjob2 variable. Hence, the fact that both parents of a family work fulltime or not play no role in their job satisfaction.

Actual values / percentage %100 of the actual values	Bothft	
Satjob2	No	Yes
Very satisfied	132/45%	74/45%
Not very satisfied	162/55%	92/55%

Table 5: Actual values and percentage %100 of the actual values of the independent non-correlated variable bothft in terms of the dependent variable satjob2

Table 6 presents analytically the total amount of values in the 4 categories of the income4 variable in terms of the categories of satjob2. The correlation of these 2 variables is shown by the fact that high wages (60.000 or more) lead to job satisfaction whereas lower wages (60.000 or less and especially 24.999 or less) leads to not very satisfied employees.

Actual values / percentage %100 of the actual values	Income4			
Satjob2	24.999 or less	25.000 to 39.000	40.000 to 59.999	60.000 or more
Very satisfied	33/30 %	59/48.4 %	48/47.5 %	66/52.8 %
Not very satisfied	79/70 %	63/51.6 %	53/52.5 %	59/47.2 %

Table 6: Actual values and percentage %100 of the actual values of the independent correlated variable income4 in terms of the dependent variable satjob2

4.2 Correlations between income4 (dependent) and the rest of variables (independent)

In this section we will examine the relationships between the dependent variable income4 and the rest of the variables which are considered as independent in this case study.
For the analysis of the correlation of the variables, we decided to work with the Kendall rank correlation(non-parametric) since our sample size is relatively small (460 observations) and the dependent variable income4 has many tied ranks. The variables are measured on an ordinal or continuous scale so we can use Kendall Rank Correlation accordingly.
On the Table 7 we display the correlation coefficient of Kendal as a result from the tau value and the p-value, between our dependent variable (income4) and the list of the independent variables:

	Positive and Statistical Important	Positive and NOT statistical important	Negative and Statistical Important	Negative and NOT statistical important
Strong (+ or – 0.30 or above)	Rincom91 bothft			
Moderate (+ or – 0.20 to 0.29)	Educ, degree			
Weak (+ or – 0.10 to 0.19)	Age, hrs1, husbhr, agecat	Sex,jobinc, impjob,husbft, wifeft,wifehr	Satjob, satjob2	

Table 7: Correlation coefficient of Kendall for dependent variable income4

As a result, we can understand that the variables rincom91 and bothft have a strong positive and statistical important correlation with our dependent variable income4. Next, the variables educ and degree have a moderate positive and statistical important correlation with the variable income4. The variables age, hrs1, husbhr and agecat4 have a weak positive and statistical important correlation with the variable income4.

Also, the variables sex, jobinc, impjob, husbft, wifeft, wifehr have a weak positive and not statistically important correlation with the dependent variable. Finally, the variables satjob, satjob2 have a weak negative and statistically important correlation with the variable income4.

For further investigation regarding the correlation between the independent variables in combination with the dependent variable income4, we decided to use the library tree. The same results as found from the Kendall's correlation are also produced though the generated tree. More specifically, as shown in the Figure 8, the variables that were actually used for the construction in tree construction are: rincom91, bothft and hrs1, with the most important ones to be rincom91 and bothft.

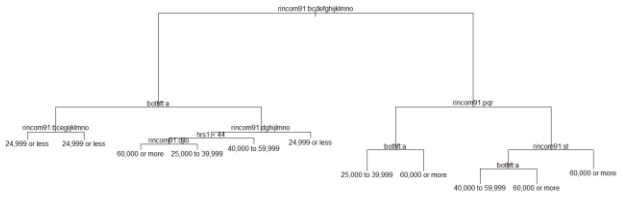

Figure 8: Tree for dependent variable income4

We also use the Chi square statistic to determine if two categorical variables have a significant dependency between them. The null hypothesis of the Chi-square test is that no relationship exists on the categorical variables in the population, so they are independent.

In the following Table 8 we have gathered the outcome of the chi-square test p-values for each of the categorical variables.

Chi-square test p-values	Income4
Sex	0.3948
Degree	**$9.1*10^{-10}$**
Satjob	0.09528
Satjob2	0.01981
Rincom91	**$2.2*10^{-16}$**
Jobinc	0.5207
Impjob	0.2443
Bothft	**$6.1*10^{-15}$**
Agecat4	**0.001236**

Table 8: Chi-square test p-values for dependent variable income4

As expected from the previous analysis, the categorical variables degree, rincom91, bothft and agecat4 have a p-value less than the significance level of 0.05, so we reject the null hypothesis and conclude that the two variables are in fact dependent.

Finally, the rest of the variables and more specifically the variables: sex, satjob, satjob2, jobinc, impjob have a p-values more than the significance level of 0.05, so we accept the null hypothesis and conclude that the two variables are independent.

4.3 Correlations between rincom91 (dependent) and the rest of variables (independent)

For our last dependent variable rincom91 we have used multiple tests to find the correct independent variables to use in our models. At first, we started with plotting the barplots of rincom91 in relation to the other categorical variables.

We can observe, in Figures 9,10 and 11, that for all datasets, variable impjob and variable degree are greatly saturated towards a specific value (Very important and High School). That can be an indicator of a bad predictor and we are going to investigate it in the next tests.

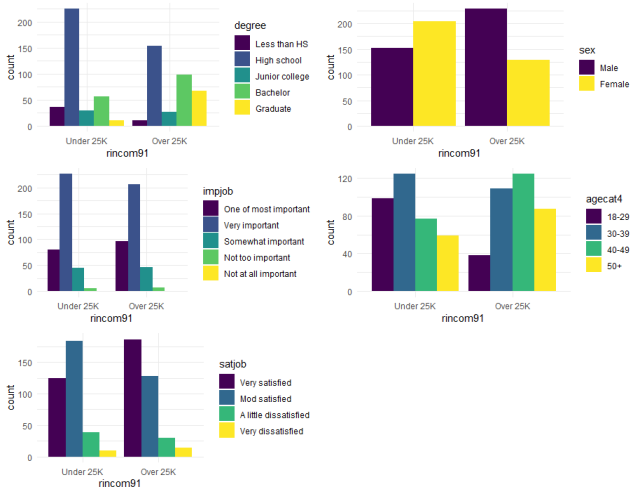

Figure 9: Barplots of dataset with two classes on rincom91

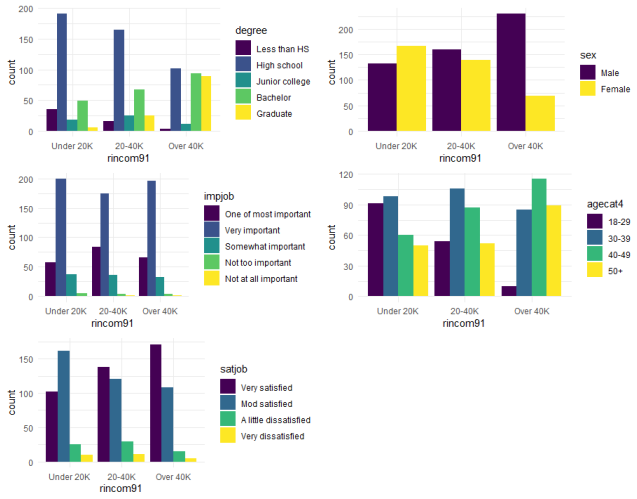

Figure 10: Barplots of dataset with three classes on rincom91

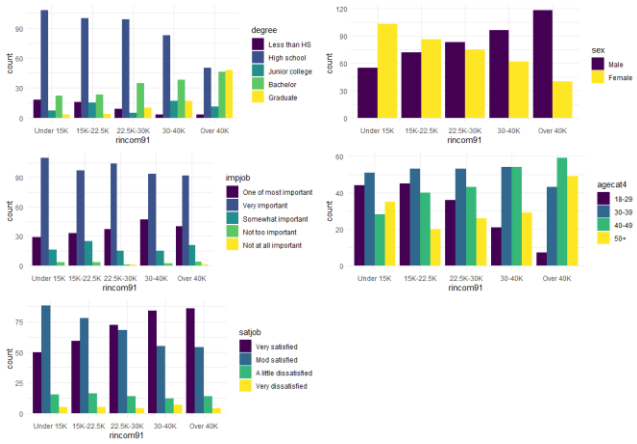

Figure 11: Barplots of dataset with five classes on rincom91

We also created barplots with quadruple correlation (rincom91, age, sex, degree). In Figures 13 and 14 we can observe that impjob does not contain enough observations for us to consider it and it will probably confuse our model. The same observation can be made for Figure 12 and variable degree. Both age and sex seem to be good predictors since all cells follow a specific pattern-like flow in all datasets.

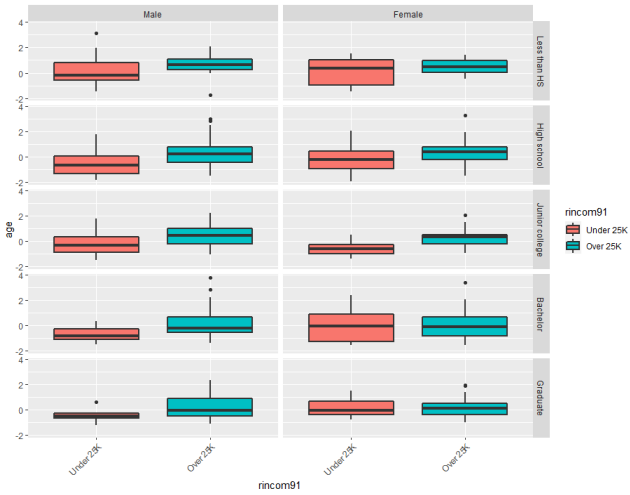

Figure 12: Boxplots of dataset with two classes on rincom91

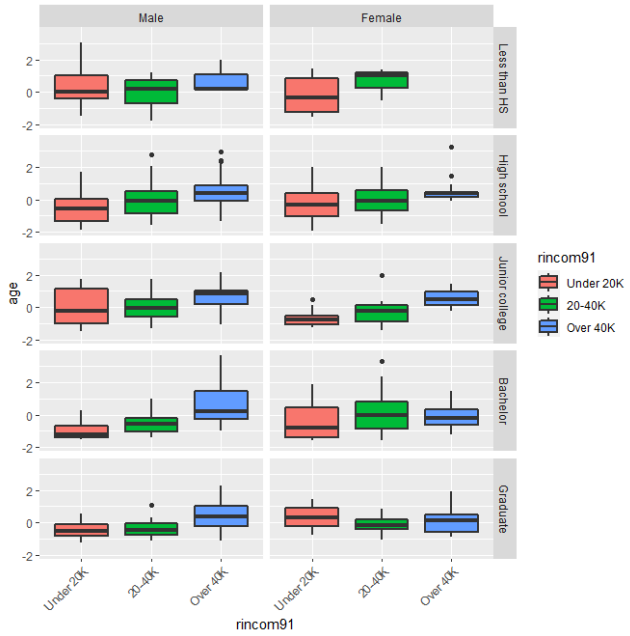

Figure 13: Boxplots of dataset with three classes on rincom91

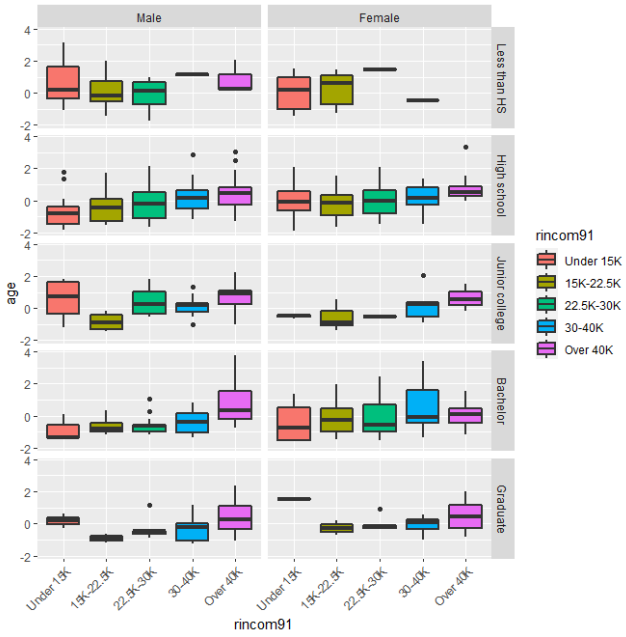

Figure 14: Boxplots of dataset with five classes on rincom91

We performed Pearson's Chi-square test to find which variables are significant towards rincom91. The results can be found in Table 9 below.

Pearson's Chi-squared test p-value	rincom91
age	**0.01268**
educ	**2.2e-16**
hrs1	**0.0001176**
sex	**6,13E-13**
degree	**2.2e-16**
satjob	**2,22E-03**
impjob	0.3741
agecat4	**2.2e-16**

Table 9: P-values of Pearson's Chi-squared test

Using the p-values we can reject the null hypothesis for all variables except impjob.

One more test that we ran is the dropterm from the MASS package. Dropterm tries fitting all models that differ from the current model by dropping a single term, maintaining marginality. Running this model gave us the following results that can be seen in Figure 15.

```
Single term deletions

Model:
rincom91 ~ sex + age + hrs1 + educ + degree + impjob + satjob
       Df    AIC    LRT   Pr(Chi)
<none>    1626.1
sex     1 1677.0 52.897 3.515e-13 ***
age     1 1712.2 88.033 < 2.2e-16 ***
hrs1    1 1647.2 23.053 1.576e-06 ***
educ    1 1625.3  1.231 0.2672525
degree  4 1648.9 30.753 3.439e-06 ***
impjob  4 1621.0  2.936 0.5686562
satjob  3 1636.9 16.758 0.0007927 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 15: Single Term deletion (dropterm) results considering rincom91 as dependent variable

The statistically important variables are now sex, age, hrs1, degree and satjob. Using both tests we accept all variables except impjob since in both our tests it is not statistically important.

5 Models

5.1 GLM modeling framework considering satjob2 as the dependent variable and the rest as independent

Considering the results in the correlation section 4.1 we gathered the most important variables for the dependent variable satjob2 i.e., the ones that had the highest correlation with it, and we proceeded to the construction of our models. More specifically, in terms of the numeric variable, although educ and age could not be labeled as correlated, we experimented on several of their possible transformations and concluded to the fact that the addition of the variables log(age) and educ^2 improves the overall model performance as shown in Figure 16. Furthermore, in terms of the categorical independent variables we added to our model the correlated ones with the satjob2 variable. We only excluded the satjob variable because it is a reformed version of the satjob2 variable, and its addition would lead to misleading and overfitting results.

Nevertheless, the results of our two best performing models are not decent enough leading to the result that although there seem to be some correlations between the independent variables and the dependent variable satjob2 the current data can't produce great results considering satjob2 as the dependent variable. We observe that we have some statistically significant terms from the jobinc, income4 and impjob variables. Also, in terms of predictions our model predicts decently the not very satisfied class with 70% accuracy whereas it does not predict decently enough the very satisfied class resulting in the lower 60% accuracy which leads us to the overall 65% mean model accuracy.

```
> #65 mean accuracy
> model1 =  glm(satjob2~jobinc+income4+impjob+rincom91+degree+I(educ^2)+log(age),binomial)
> summary(model1)

call:
glm(formula = satjob2 ~ jobinc + income4 + impjob + rincom91 +
    degree + I(educ^2) + log(age), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0011  -1.1023   0.7023   1.0342   1.6875

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  2.2070139  1.4572933   1.514  0.12991
jobincSECOND                -0.5272296  0.2950392  -1.787  0.07394 .
jobincTHIRD                 -0.6972194  0.2891207  -2.412  0.01589 *
jobincFOURTH                -0.9125144  0.3296427  -2.768  0.00564 **
jobincFIFTH                 -0.4696484  0.5892929  -0.797  0.42547
income425,000 to 39,999     -0.8180656  0.2871551  -2.849  0.00439 **
income440,000 to 59,999     -0.5889720  0.3200831  -1.840  0.06576 .
income460,000 or more       -0.7158368  0.3476101  -2.059  0.03946 *
impjobVery important         0.1977617  0.2390596   0.827  0.40810
impjobSomewhat important     1.2101117  0.3886265   3.114  0.00185 **
impjobNot too important      0.7490339  0.7631654   0.981  0.32635
impjobNot at all important   0.4273865  1.5006302   0.285  0.77579
rincom91$60000 or more      -0.3625657  0.2925454  -1.239  0.21522
degreeGraduate              -0.2766836  0.4071048  -0.680  0.49673
I(educ^2)                    0.0005715  0.0017805   0.321  0.74821
log(age)                    -0.3187588  0.3679636  -0.866  0.38634
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 632.68  on 459  degrees of freedom
Residual deviance: 591.61  on 444  degrees of freedom
AIC: 623.61

Number of Fisher Scoring iterations: 4

> probs = predict(model1,type ="response")
> pred.classes = ifelse(probs > 0.5, "pos", "neg")
> prop.table(table(satjob2,pred.classes),1)
                  pred.classes
satjob2                    neg       pos
  Very satisfied     0.5970874 0.4029126
  Not very satisfied 0.2992126 0.7007874
```

Figure 16: Best performing model in terms of accuracy considering satjob2 as the dependent variable and the rest as independent.

In Figure 17, we present our second-best performing model in terms of accuracy which results in 62% mean accuracy, 61% for the very satisfied class and 63% for the not very satisfied class. Here, both of our 3 independent variables namely jobnic, income4, and impjob are statistically significant terms whereas in the previous model we had 4 of the total 7 of the independent variables which were considered as not statistically important from our model namely rincom91, the logarithmic transformation of the age variable, the square of the educ variable and the degree variable. Nevertheless, their inclusion to the model improved it by 3% (from 62% to 65% mean accuracy). On the other hand, it transformed our model to a more complex and harder to interpret model adding to its overall complexity. Both best performing models prove the fact that the current data don't provide enough trends between the independent variables and the dependent satjob2 variable to build a proper performing model.

```
> #65 mean accuracy
> model1 =  glm(satjob2~jobinc+income4+impjob,binomial)
> summary(model1)

Call:
glm(formula = satjob2 ~ jobinc + income4 + impjob, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0215  -1.1058   0.6723   1.0335   1.5242

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                  1.1648     0.3508   3.320 0.000899 ***
jobincSECOND                -0.5524     0.2913  -1.897 0.057879 .
jobincTHIRD                 -0.7050     0.2872  -2.455 0.014097 *
jobincFOURTH                -0.9271     0.3276  -2.830 0.004656 **
jobincFIFTH                 -0.5033     0.5856  -0.860 0.390033
income425,000 to 39,999     -0.8379     0.2830  -2.960 0.003073 **
income440,000 to 59,999     -0.7803     0.2949  -2.646 0.008145 **
income460,000 or more       -1.0239     0.2861  -3.578 0.000346 ***
impjobVery important         0.2073     0.2372   0.874 0.382137
impjobSomewhat important     1.2429     0.3849   3.229 0.001241 **
impjobNot too important      0.7350     0.7587   0.969 0.332657
impjobNot at all important   0.3414     1.4686   0.232 0.816192
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 632.68  on 459  degrees of freedom
Residual deviance: 595.13  on 448  degrees of freedom
AIC: 619.13

Number of Fisher Scoring iterations: 4

> probs = predict(model1,type ="response")
> pred.classes = ifelse(probs > 0.5, "pos", "neg")
> prop.table(table(satjob2,pred.classes),1)
                  pred.classes
satjob2                    neg       pos
  Very satisfied     0.6116505 0.3883495
  Not very satisfied 0.3661417 0.6338583
```

Figure 17: Second best performing model in terms of accuracy considering satjob2 as the dependent variable and the rest as independent

5.2 Ordinal Logistic Regression modeling framework considering income4 as the dependent variable and the rest as independent

The dataset has a dependent variable known as income4. It has 4 levels namely "24,999 or less", "25,000 to 39,999", "40,000 to 59,999", "60,000 or more". This situation is best for using ordinal regression because of presence of ordered categories. For building this model, we will be using the polr command to estimate an ordered logistic regression. Then, we'll specify Hess = TRUE to let the model output show the observed information matrix (Hessian) from optimization which is used to get standard errors.

First of all, for the construction of our model we experimented using our dependent variable, income4, in combination with different variations of our independent variables focusing on the ones with the most correlation.

In our results, we see the usual regression output coefficient table including the value of each coefficient, standard errors, t values, estimates for the two intercepts, residual deviance, and AIC. AIC is the information criteria. Lesser the better. Also, to understand better our model we calculate some essential metrics such as p-value, confidence intervals, odds ratio. The results can be found in the Figure 18 below.

```
> model <-polr(income4~educ+bothft+hrs1+ï..age+as.numeric(rincom91), Hess=TRUE, data = df1)
> summary(model)
Call:
polr(formula = income4 ~ educ + bothft + hrs1 + ï..age + as.numeric(rincom91),
    data = df1, Hess = TRUE)

Coefficients:
                     Value Std. Error t value
educ                0.130352   0.039473  3.3023
bothftYes           2.340523   0.226055 10.3538
hrs1                0.008476   0.008626  0.9827
ï..age              0.014303   0.009095  1.5727
as.numeric(rincom91) 0.460216   0.037854 12.1576

Intercepts:
                                        Value   Std. Error t value
24,999 or less|25,000 to 39,999          8.7394  0.8809     9.9210
25,000 to 39,999|40,000 to 59,999 10.9552  0.9392    11.6643
40,000 to 59,999|60,000 or more   12.6816  0.9909    12.7978

Residual Deviance: 890.4218
AIC: 906.4218
```

Figure 18: Summary of the model produced with the Ordinal Logistic Regression modeling framework

The interpretation for the coefficients is as follows. For example, holding everything constant, an increase in value of educ increase the expected value of income4 by 0.130352 and an increase in bothft generates an increase of the expected value by 2.340523. Likewise, the coefficients for the rest can be interpreted. Note that the ordinal logistic regression outputs multiple values of intercepts depending on the levels of intercept. For example, the "24,999 or less" | "25,000 to 39,999" intercept takes value of 8,7394 indicating that the expected odds of identifying in "24,999 or less" category, when other variables assume a value of zero, is 8,7394. Moreover, the p-value for all the independent variables was less than 0.05, so we can say that they have a significant role in the interpretation of the dependent variable income4.

After building and interpreting the model, the next step is to evaluate it. A basic evaluation approach that we used is to compute the confusion matrix and the misclassification error.

The confusion matrix, in Figure 19, shows the performance of the ordinal logistic regression model. For example, it shows that, 78 times the category "24,999 or less" is identified correctly or that 80 times the category "25,000 to 39,999" is identified correctly. To produce the final results, we used the stepAIC function in order to include the best performing independent variables in our model. On the initial models our misclassification error rate was between 50% to 60%. After the use of stepAIC we managed to improve our model and we find that the misclassification error for our final model is 35%.

```
> predict_income = predict(model,df1)
> table(df1$income4, predict_income)
                  predict_income
                   24,999 or less 25,000 to 39,999 40,000 to 59,999 60,000 or more
  24,999 or less               78               33               1              0
  25,000 to 39,999             18               80              19              5
  40,000 to 59,999              5               23              45             28
  60,000 or more                6               16               7             96
> mean((df1$income4) != (predict_income))
[1] 0.35
```

Figure 19: Prediction table and misclassification error of the model

Because the interpretation of the logistic ordinal regression in terms of log odds ratio is not easy to understand, we also worked on a different approach for the interpretation using plots and more specifically we used the "effects" library to generate the plots that we display next.

For example, the Figure 20 shows the effect of the independent variable bothft on the dependent variable income4. More specifically, it shows that the variable bothft increases the likelihood of classification in "40,000 to 59,999" and "60,000 or more" classes, while decreasing the likelihood of classification in "24,999 or less" and "25,000 to 39,999".

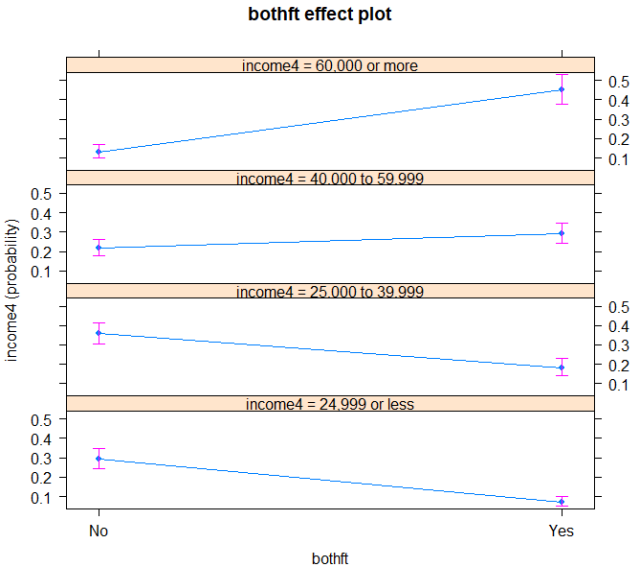

Figure 20: Effect of bothft on income4

It is also possible to look at the joint effect of two independent variables. On the Figure 21 below, we display the joint effect of variables bothft and age.

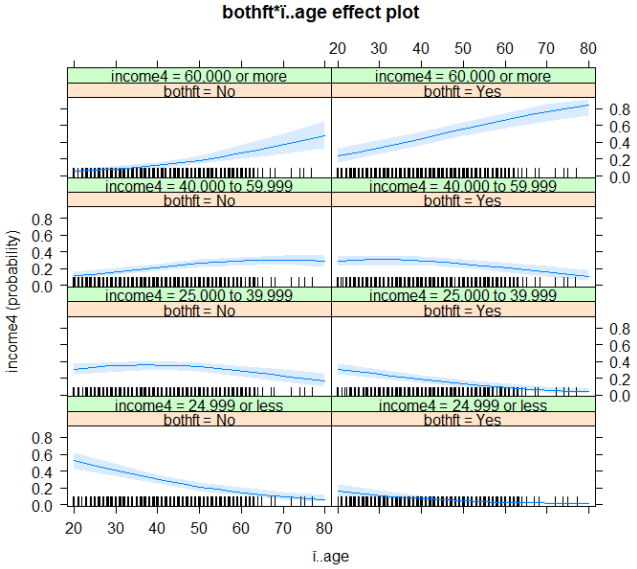

Figure 21: Joint effect of bothft and age on income4

Observing for example, on the top row, we notice that the interaction of bothft and age increases the likelihood in category "60,000 or more" or in the bottom row it decreases the likelihood in category "24,999 or less".

On the Figure 22, we display the predicted probabilities of each outcome for the independent variables bothft and age in combination with our dependent variable income4.

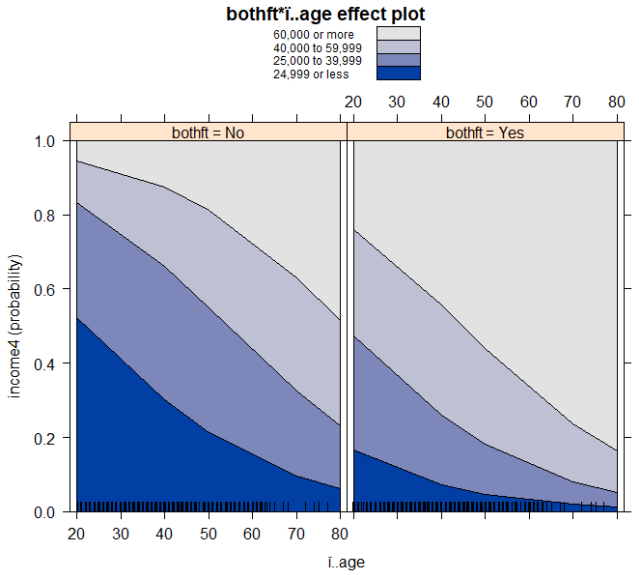

Figure 22: Plot predicted probabilities of each outcome

We notice that when the value of bothft is negative there is a higher percentage of probability for the classes "24,999 or more", "25,000 to 39,999" and "40,000 to 59,999" in comparison to the positive one. Finally, when the bothft value is "Yes" then we notice the opposite, meaning that the percentage is higher for "60,000 or more" instead of the value being "No".

5.3 Models considering rincom91 as the dependent variable and the rest as independent

For our last dependent variable, we experimented with four different algorithms (Polr, Multinom, RandomForest (ranger) and XGBoost). As we have previously stated in section 4.3 the independent variables that were used were age, educ, hrs1, sex, degree, satjob and agecat4. For all our models we scaled the numeric variables (age, educ, hrs1) using R's scale() function. Moreover, SMOTE was used to oversample the minority classes, so that every class contains the same number of observations. Since rincom91 originally contains eighteen "classes" we aggregated these classes into fewer new classes and extracted three different datasets. The aggregation of our classes can be found in Table 10.

Dataset	Classes	#Classes
A	Under 25K Over 25K	2
B	Under 20K 20-40K Over 40K	3
C	Under 15K 15K-22.5K 22.5K-30K 30-40K Over 40K	5

Table 10: The three newly created datasets

Since Polr needs at least three classes we ran the algorithm only for Datasets B and C. For all four models, we used a 10-fold cross validation (from the caret package) to find the best parameters for each algorithm. The metric that we used in the cross validation was Accuracy. Finally, we split the datasets into train-test splits with a ration 80:20. The train split was used to train the dataset and the test split to predict the classes. On sections 5.3.1 to 5.3.4 we are going to focus solely on Dataset B for simplicity.

5.3.1 Polr (Ordered Logistic or Probit Regression)

The first model we built for the dependent variable rincom91 was Probit Regression. After training our model with the cross-validation technique we ended up with the following results:

```
Confusion Matrix and Statistics

          Reference
Prediction  Under 20K 20-40K Over 40K
  Under 20K        35     17        1
  20-40K           19     23       10
  Over 40K          6     20       49

Overall Statistics

               Accuracy : 0.5944
                 95% CI : (0.5189, 0.6669)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 6.823e-13

                  Kappa : 0.3917

 Mcnemar's Test P-Value : 0.07139

Statistics by Class:

                     Class: Under 20K Class: 20-40K Class: Over 40K
Sensitivity                    0.5833        0.3833           0.8167
Specificity                    0.8500        0.7583           0.7833
Pos Pred Value                 0.6604        0.4423           0.6533
Neg Pred Value                 0.8031        0.7109           0.8952
Prevalence                     0.3333        0.3333           0.3333
Detection Rate                 0.1944        0.1278           0.2722
Detection Prevalence           0.2944        0.2889           0.4167
Balanced Accuracy              0.7167        0.5708           0.8000
```

Figure 23: Polr Confusion Matrix and statistics.

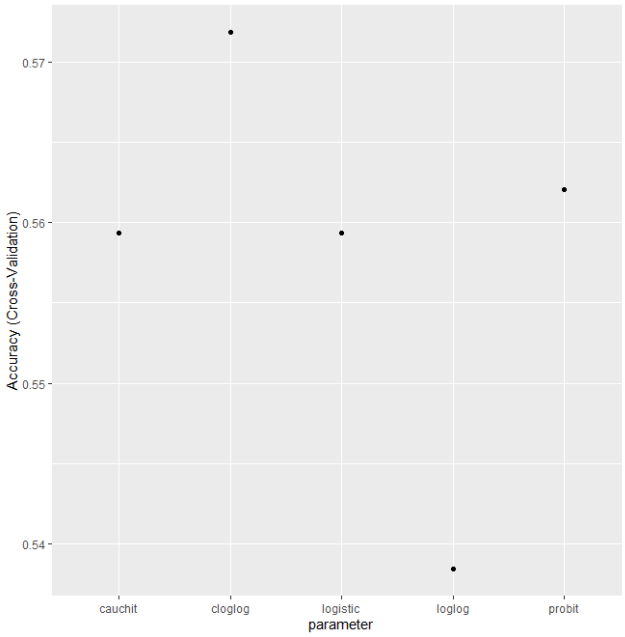

Figure 24: Polr parameters.

We observe that the accuracy of the model is 59%. That is low and can be widely interpreted by the fact that the model predicts poorly observations of classes "Under 20K" and "20-40K". Kappa value 0.39 can be considered as fair, but can be greatly improved. From figure 24 we see that the algorithm that give us the best accuracy is cloglog.

5.3.2 Multinom (Multinomial Logistic Regression)

For our next model we used the multinom function of the nnet package. The results of this algorithm can be seen on Figure 25.

```
Confusion Matrix and Statistics

          Reference
Prediction  Under 20K 20-40K Over 40K
  Under 20K        42     27        2
  20-40K           12     10        9
  Over 40K          6     23       49

Overall Statistics

               Accuracy : 0.5611
                 95% CI : (0.4853, 0.6348)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 3.059e-10

                  Kappa : 0.3417

 Mcnemar's Test P-Value : 0.003053

Statistics by Class:

                     Class: Under 20K Class: 20-40K Class: Over 40K
Sensitivity                    0.7000       0.16667          0.8167
Specificity                    0.7583       0.82500          0.7583
Pos Pred Value                 0.5915       0.32258          0.6282
Neg Pred Value                 0.8349       0.66443          0.8922
Prevalence                     0.3333       0.33333          0.3333
Detection Rate                 0.2333       0.05556          0.2722
Detection Prevalence           0.3944       0.17222          0.4333
Balanced Accuracy              0.7292       0.49583          0.7875
```

Figure 25: Multinom Confusion Matrix and statistics.

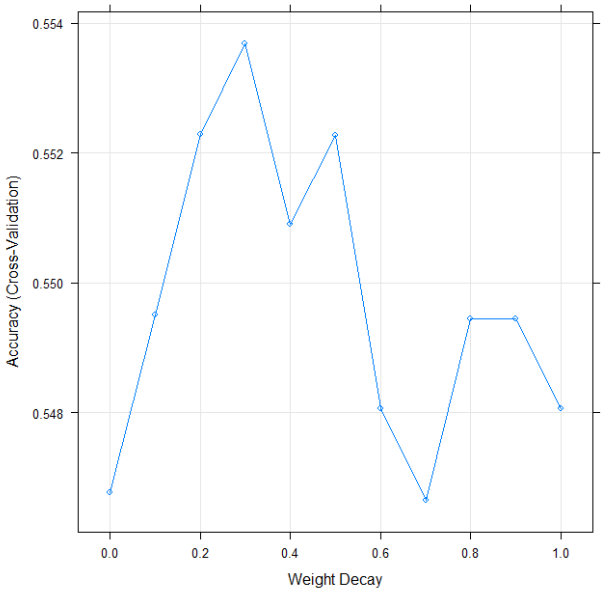

Figure 26: Multinom parameters.

Comparing our results to the previous model, we can observe straight away that this model is worse on almost every aspect. The only noticeably differences from the first model are the better prediction of class "Under 20K" and the much worse predictions of class "20-40K". Weight Decay maximizes our results in 0.3 and decreases right after as seen on Figure 25.

5.3.3 Random Forest

For our third model we decided to turn to Decision Trees and more specifically the ranger algorithm, a fast implementation of random forests. The results of this model can be viewed below.

```
Confusion Matrix and Statistics

          Reference
Prediction  Under 20K 20-40K Over 40K
  Under 20K        41     17        2
  20-40K           15     30       10
  Over 40K          4     13       48

Overall Statistics

               Accuracy : 0.6611
                 95% CI : (0.587, 0.7299)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.4917

 Mcnemar's Test P-Value : 0.7571

Statistics by Class:

                     Class: Under 20K Class: 20-40K Class: Over 40K
Sensitivity                    0.6833       0.5000          0.8000
Specificity                    0.8417       0.7917          0.8583
Pos Pred Value                 0.6833       0.5455          0.7385
Neg Pred Value                 0.8417       0.7600          0.8957
Prevalence                     0.3333       0.3333          0.3333
Detection Rate                 0.2278       0.1667          0.2667
Detection Prevalence           0.3333       0.3056          0.3611
Balanced Accuracy              0.7625       0.6458          0.8292
```

Figure 27: Random Forest Confusion Matrix and statistics

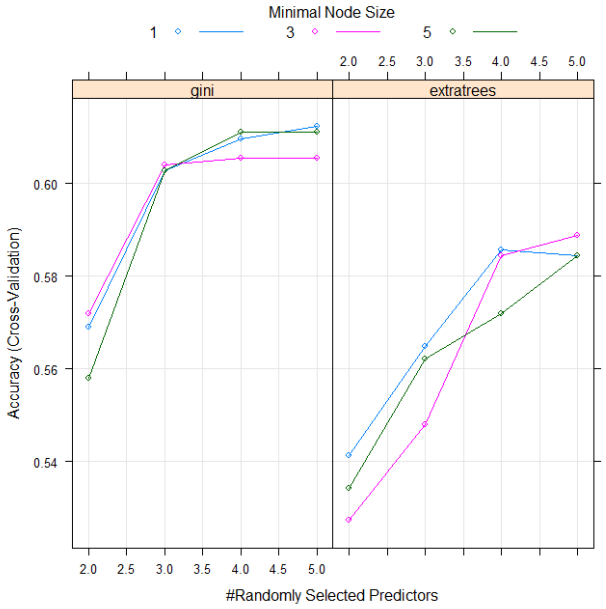

Figure 28: Random Forest parameters.

In comparison to the two previous models, we can easily observe the big improvements. Accuracy has been increased to 66% compared to the previous best 59% and the Kappa value to 0.49 considered as moderate. Apart from the great improvements, we started to notice a pattern for class "20-40K". The gini algorithm outperformed extratrees on every aspect and the best results are obtained with minimal node size = 1.

5.3.4 XGBoost

Our last model for the dependent variable rincom91 was built with XGBoost.

```
Confusion Matrix and Statistics

          Reference
Prediction  Under 20K 20-40K Over 40K
  Under 20K        37     20        5
  20-40K           19     28       11
  Over 40K          4     12       44

Overall Statistics

               Accuracy : 0.6056
                 95% CI : (0.5301, 0.6775)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 7.45e-14

                  Kappa : 0.4083

 Mcnemar's Test P-Value : 0.9807

Statistics by Class:

                     Class: Under 20K Class: 20-40K Class: Over 40K
Sensitivity                    0.6167        0.4667          0.7333
Specificity                    0.7917        0.7500          0.8667
Pos Pred Value                 0.5968        0.4828          0.7333
Neg Pred Value                 0.8051        0.7377          0.8667
Prevalence                     0.3333        0.3333          0.3333
Detection Rate                 0.2056        0.1556          0.2444
Detection Prevalence           0.3444        0.3222          0.3333
Balanced Accuracy              0.7042        0.6083          0.8000
```

Figure 29: XGBoost Confusion Matrix and statistics.

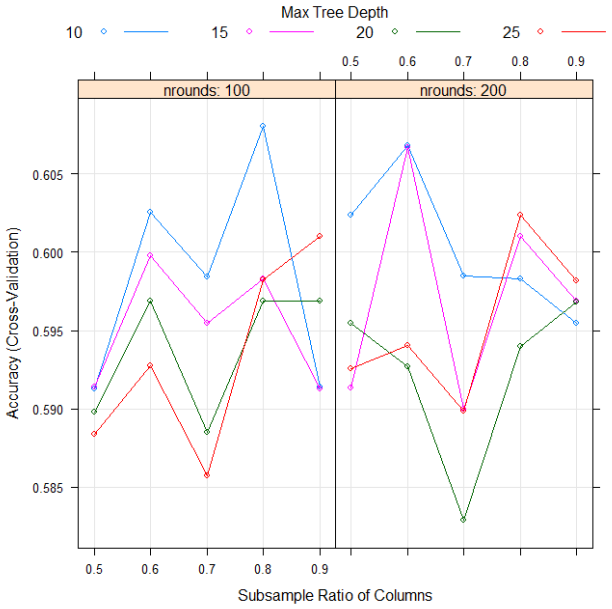

Figure 30: XGBoost parameters.

We can observe that this model performs better than the first two but a bit worse than the Random Forest model. Accuracy is 60% and Kappa value is 0.4 (moderate). Again, in this model it is clear, that we have to do some optimizations in order to avoid the misclassification error of class "20-40K". Smaller tree depths outperformed the bigger ones, and our best accuracy was obtained with 100 rounds of boosting.

5.3.5 Final results of models considering rincom91 as the dependent variable and the rest as independent.

To sum up the previous sections 5.3.1-5.3.4 we present all our results in the graphs below and conclude the paper in section 6.

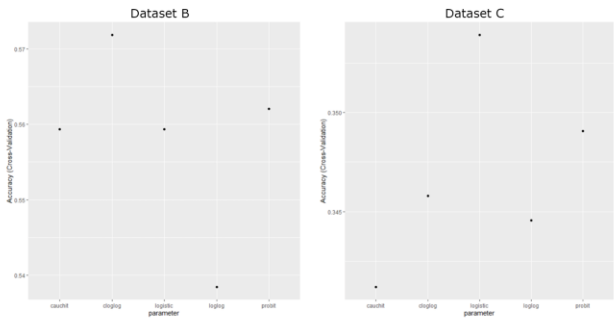

Figure 31: Polr parameters for Datasets B and C.

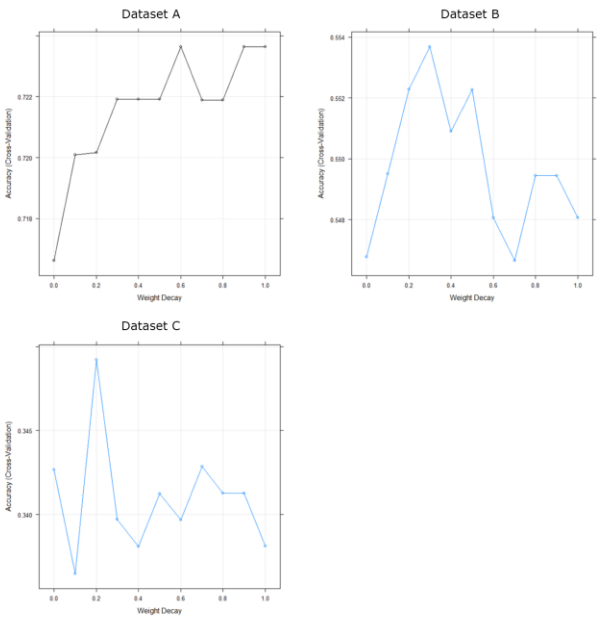

Figure 32: Multinom parameters for Datasets A,B and C.

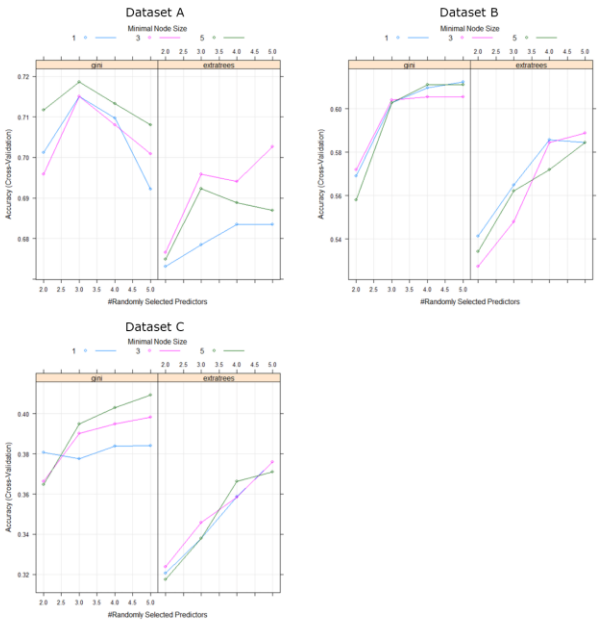

Figure 33: Random Forest parameters for Datasets A,B and C.

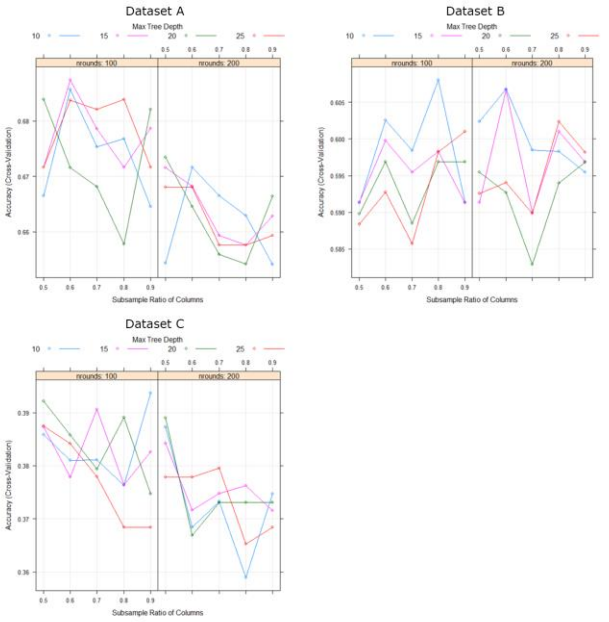

Figure 34: Polr parameters for Datasets B and C.

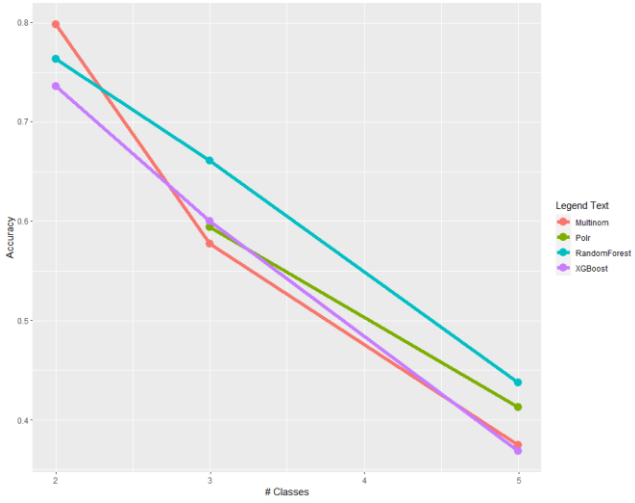

Figure 35: Accuracy of all models in Datasets A,B and C with 2,3 and 5 classes of rincom91.

6 Conclusion

Completing our research, we reached the following conclusions:

For the dependent variable satjob2 we first examined its' correlations with the independent variables to select which of them we should embed in our modeling framework. Then we created two models and each of them has a different interpretation. The first one achieves the highest accuracy i.e., 65% but it's quite complex whereas the second model is

simpler and easier to interpret having all its' terms statistically important but achieves lower accuracy i.e., 62%.

From the dependent variable's, income4, perspective, we investigated the relationships between this variable and the rest of the independent variables. We managed to separate them into independent variables with strong, mediate, and weak correlation and used them for the construction of our initial and final model. Because the variable income4, is divided into 4 levels, we used the Ordinal Logistic Regression for the model. As a result, we can say that the produced model achieves mediocre results with a misclassification error rate of 35% and that the independent variables increase the likelihood of the top 2 categories in comparison to the bottom 2 that the likelihood is decreased.

For the dependent variable rincom91, from Figure 35 we observe that the best algorithm for Dataset A is Multinom while RandomForest (ranger) is the best algorithm for Dataset B and C. The performance dropped dramatically going from Dataset A to B and C. Even though Multinom was the best performing algorithm on Dataset A it performed poorly on Dataset B and C trailing behind the rest. The Ordinal Regression Algorithm performs on par with Random Forest and XGBoost on Dataset B and starts to outperform them with 5 or more classes. Overall, the number of Classes has a big impact on every algorithm's performance and hinders their ability to predict these classes correctly.