2021-08-08
Cameron Stewart, Michael Mazel, Rick Fontenot and Tricia Herrera

Predicting Income

Introduction

Does marital status affect income? What is the likelihood that a 32-year-old white married male will make more than $50K a year compared to a 59-year-old black divorced female? An individual's income may be influenced by certain factors, and this study examines those factors individually and collectively to determine whether they do have an effect. Using four statistical models, we predict whether an individual will make more or less than $50K a year in the US in 1994. By using each model's AUC against the test set, we can determine if one model methodology outperforms another in predicting income. The models will select a cutoff that balances accuracy, sensitivity, and specificity.

Data Description

This study used data from Barry Becker's 1994 Census database. In the data set, 32,561 observations are included along with 14 attributes. There are 6 continuous variables and 8 categorical attributes in this data set. A copy of it is available in the machine learning repository at UCI [1]. Detailed information about each attribute is available in the Appendix, Data Description.

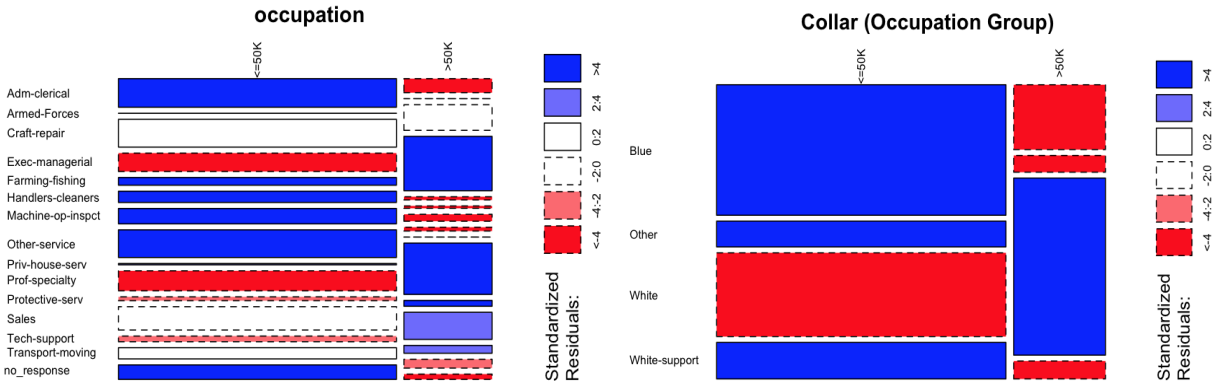
Exploratory Analysis

Upon initial examination of the data available we found that 5.6% observations are missing values for both occupation & workclass and 1.7% are missing native country information. Through discussion we decided a missing value on occupation and workclass could be a non-response indicating their occupation could be hard to nail down if the person works multiple low wage jobs simultaneously to make ends meet or if they are in a transition between occupations. This lack of response could be a useful predictor, so rather than drop observations with missing values, we replaced with "no_response". See appendix figure EDA1.

For this population, only 24% of the people earn more than $50K. In reviewing summary statistics, we found many categorical variables to be heavily imbalanced, for instance US accounts for 90% of the native country observations and the private sector accounts for 70% of the working class. Numerical variables have relatively normal distributions with age truncated on the left due to working age requirements, and capital gains/losses primarily zero with a few observations in long right skews. For more information, see appendix figures EDA2 and EDA3.

For each categorical variable, we studied significant differences in income by level and whether groupings and level-reduction would improve modeling and simplify interpretation. For example, occupation was regrouped to office work (white-collar), physical labor intensive (blue collar) or other (primarily the no_response observations). Within the white-collar jobs, Admin-

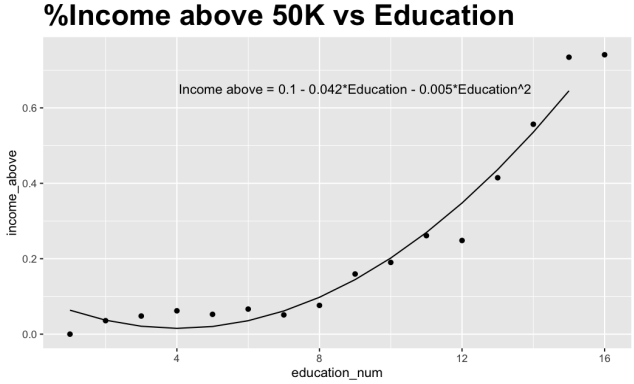
clerical had a different relationship to income than managerial and technical occupations, so it was set into its own level for support role within white collar jobs (new variable collar will now be referenced). White Collar jobs appears to have one of the strongest differences between income groups with 40% earning above $50K and other groups only having 13%-17% above $50K.



These variables also benefited with level groupings/reduction for model performance improvements and simpler logistic regression interpretation.

- Workclass: Regrouping to Private, Government, Self Employed, Not Working, and Other (new variable work_sector will now be referenced) See appendix figure EDA4
- Marital Status: Regrouping to just Married, Previously Married, and Single (new variable marriage_status will now be referenced) See appendix figure EDA5
- Education: When levels are re-ordered this correlate perfectly with education_num and the numerical version of this variable shows better correlation to income, see next section for more information. See appendix figure EDA7

For each numerical variable the log-odds of people with income above $50K was calculated and studied for linear vs. nonlinear relationships. For example, Education: Regression with Education^2 improves adjusted R^2 from 0.79 to to 0.96 (for simple interpretability the proportion is plotted below in lieu of log-odds), and this fit curve with education appears to be one of the strongest predictors of income; 72% of people with advanced degrees earning above $50K, and less than 18% for those with high school graduation or less.



The following variables also exhibit non-linear relationships and quadratic terms produced better fits that will be used in more complex models:

- Age: Regression with Age^2 and Age^3 improves adjusted R^2 from 0.0 to to 0.78, see appendix figure EDA9
- Hours per Week: Regression with $Hours^2$ improves adjusted R^2 from 0.15 to 0.24. Since the fit is still not ideal, this variable may provide better predictions with non-parametric modeling. Eee appendix figure EDA11

The Fnlwgt variable, which stands for final weight, represents how much of a state's population may be like the demographics of each observation. The scale of this weighting varies by state and we do not have location information to do any normalization on ranges. This information is not a predictor of that specific observation's income, so we have dropped it from predictive modeling.

Objective 1

Our team set out to create a simplified interpretable model to understand the features that explain whether someone makes over $50K in 1994 in the US. To do this, our team created an 80/20 training and test split of the original data. Next, we used the EDA to guide the initial model creation and refined the model using significance testing and AIC. After selecting a model, we verified the assumptions were met and interpreted the selected predictors.

1.1 Model Selection

A base logistic regression model using all original available variables was built to serve as a benchmark and initial exploration of variable effects. Using the EDA, we created reduced categorical and eliminated fnlwgt and education from the initial model.Upon running the initial simplified model, removing the Native Country and Work Sector variables resulted in a lower AIC. The remaining variables have strong significance and any of their removal would result in a higher AIC. Our final simplified model includes age, education_num, race, sex, captial_gain, capital_loss, hours_per_week, marriage_status, and collar. Summary of coefficients can be found in Appendix - Figure SLR1.

1.2 Checking Assumptions

Assumptions of logistic regression: observations are independent, for interpretation, explanatory variables should have little to no correlation, model is sensitive to outliers. We will assume the observations are independent based on the study context. We also confirmed there is no evidence of multicollinearity based on the VIF values being far below the threshold of 10 where we would be concerned. VIF values for the model can be found in Appendix - Figure SLR2.

1.2.1 Influential point analysis (Cook's D and Leverage)

To determine if there are any influential points, we must look at the magnitude of the residual and leverage of the observations. The metric we used to consider both residual and leverage to determine influence is Cook's D. Measuring Cook's D across all observations in the training set (Appendix - Figure SLR3), we can see that observation 20872 is the most influential point by a

large margin relative to the other observations. Comparing the coefficient estimates in the original model without observation 20872 (Model 1) to the original Model with all observations (Model 2) in Appendix - Figure SLR4, we can see that the influential point removal has a minimal impact on the model coefficients. Therefore, we will proceed with all observations in the model. Appendix-Additional Influential Point Analysis Details provides a more detailed look at the magnitudes of the residual and leverage of observations.

1.3 Parameter Interpretation

After transforming our predictor estimates (e^(coefficient estimate)), we can interpret the odds ratio of our predictors and view their corresponding 95% confidence intervals in Appendix - Figure SLR7.

1.3.1 Interpretation

Below, we list the interpretation for one categorical predictor (collar) and one continuous predictor (education_num) as an example. Interpretations for all predictors can be found in Appendix - Figure SLR8.

We will assume the interpretations below are valid only when all other predictors are held constant:

Sample Interpretations:

- Collar - the odds of earning over $50K for the White-Collar population is expected to be 125.35% higher than the Blue-Collar population
- Education_num - for every 1-year increase in education, we expect the odds of earning over $50K to increase by 33.71%

1.3.2 Confidence Intervals

The 95% confidence interval for the odds ratio of each predictor is shown in Appendix - Figure SLR7. These represent the upper and lower bounds of the expected odds ratio for each predictor. If the confidence interval contains 1, then we can assume the numerical predictor or categorical level compared to the reference is not significant.

1.4 Final conclusions

Using our EDA and manual refinement based on AIC, we were able to generate a simplified logistic model. We walked through the assumptions and interpretations of the model. Looking at examples like education_num and collar, we were able to see the clear visual trends with the income variable uncovered in the EDA translate into statistically significant and interpretable odds ratios in the model. Overall, the model gave us significant insight into how all our selected predictors influence the odds of earning over $50K. Our simple logistic model has an AUC of 0.896 (Appendix - Figure SLR9 & SLR10).

Objective 2

2.1 Model Selection

A total of three models were built to compete against the simplified logistic regression model. Since the focus was on prediction rather than interpretation, a complex logistic regression model, a quadratic discriminant analysis (QDA) model, and a random forest model were developed.

Complex Logistic Regression

Using insights from our exploratory analysis, the quadratic terms Age^2, Age^3, $Education^2$, and $Hours^2$ were added to the model. A logistic regression model including all possible interaction terms was run to search for candidates, and the top ones were explored but ultimately no interaction terms were found that improved the model, so they were not used.

With quadratic terms included, AUC-test improved significantly compared to the simple model. Age^2 was not significant in the logistic regression and was dropped, AUC remained the same without Age^2 and all remaining variables were significant. Our complex logistic model has an AUC of 0.902. See Appendix - Figure CLR1 & CLR2

Quadratic Discriminant Analysis

To improve on prediction performance metrics, robust classification methods such as LDA and QDA were considered. To prepare our LDA or QDA model, we created a data frame using the following continuous variables: age, education_num, capital_gain, capital_loss and hours_per_week.

Ideally, for an LDA or QDA analysis, the assumptions that must be met are: the predictors of each response category must follow a multivariate normal distribution, constant variance, independence, and lastly, an identical covariance matrices for LDA and for QDA, each having its own covariance matrix. To verify our assumption of homogeneity of the covariance matrices, we performed a Box's M test. See Appendix - Figure QDA2 The results had a p-value of 2.2e-16, thus we reject the null hypothesis. There is evidence the covariance matrices are not equal across all groups. With this violation, QDA is more appropriate than LDA modeling. A Shaprio-Wilk test was performed to check the multivariate normality assumption. See Appendix - Figure QDA1 With a p value of 2.2e-16 there is evidence to suggest we do not have multivariate normality. In spite of applying multiple meaningful feature reduction techniques and transformations to improve the multivariate normality, no improvements to meet normality were obtained; therefore, we will proceed with caution. We will assume the observations are independent based on the study context.

To construct our QDA model, we divided our data frame of continuous predictors into an 80/20 test and train split. Our training data was fitted with a QDA model using the built-in QDA function in R. We then applied the prediction function to our test data to predict values based on the QDA model. Our QDA model has an AUC of 0.825. See Appendix - Figure QDA3 & QDA4

Ideally, all independent variables in LDA should be normal, but this assumption is debatable in our case with heavy tails for variables such as age, capital gain and capital loss. QDA does not assume classes are equal in size; however, when classes are unequal, QDA performance may be biased by class size.

Random Forest

All the models mentioned until this point had parametric assumptions, and therefore, they relied on a particular distribution of the data. Random forest was chosen as an alternative, nonparametric method in predicting income. Unlike regression models, random forest naturally captures polynomial and interaction terms, and thus they need not be included. When inputting variables into a random forest, the only preliminary steps involve transforming categorical types.

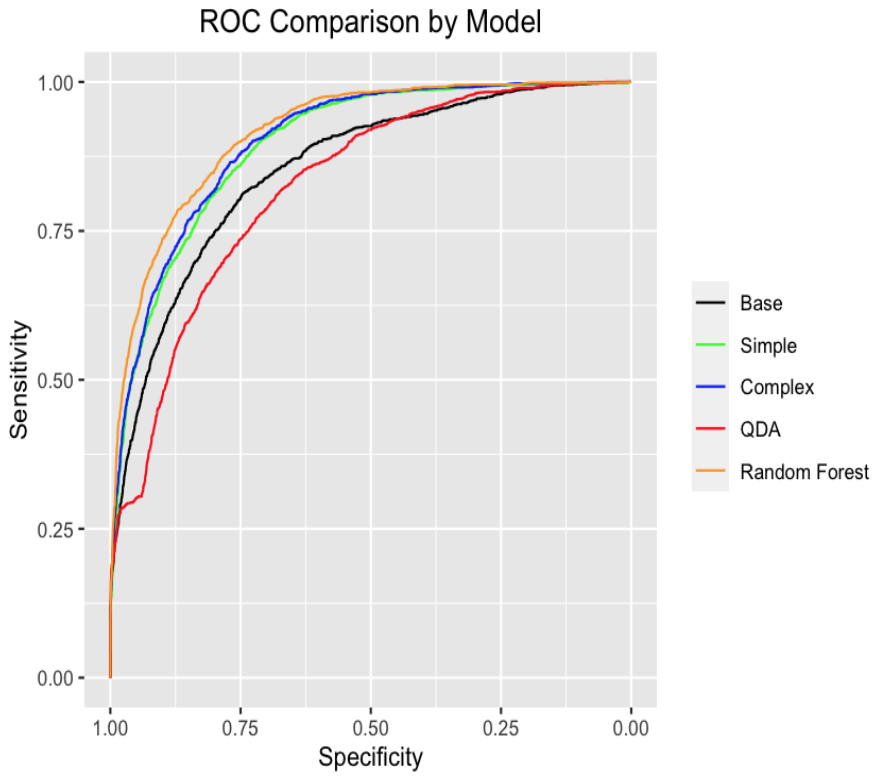
Different models considered for random forest include those of quantitative type, a mix of quantitative and categorical using the caret library's default one hot encoding, and a mix using target encoding. The "ranger" algorithm, an extremely randomized variant of random forest, was considered as well. This allowed for different split rules to occur, such as "gini" and "extratrees." Various iterations of these models took place to tune the optimal splitrule or mtry parameter. Specifically, cross validation on the training set was used to test for the optimal hyperparameter across folds. The default number of trees, 500, was found to be sufficient in generating reliable results.

The model that performed best on our test set was the traditional random forest with quantitative variables and categorical variables handled with caret's default method. With a mtry of 2, this random forest model produced an AUC of 0.918. See Appendix - Figure RF2 & RF3.

2.2 Main Analysis Content

The ROC plot below illustrates the performance of the various models implemented throughout this analysis. Random forest has the highest AUC, with the complex and simple logistic regression closely behind. The QDA model, having the lowest AUC, then followed. With random forest's ability to minimize overfitting via ensembling, the model was able to produce the highest AUC and accuracy scores. Random forest also likely benefited from the nonparametric requirement. As revealed in the EDA, hours per week appeared to improve as more polynomial terms were added. Random forest is quite effective in capturing these complex relationships. If the goal is maximizing balanced accuracy, then this model could be seen as the preferred choice. However, other models offer clearer interpretability, which may make it the preferred option. The simple logistic regression model generated an AUC score only .02 away from random forest. In many circumstances, this could be more useful due to it being more interpretable than random forest. Because logistic regression is derived from probabilities, it can easily explain the impact of a predictor variable on income in terms of odds.

For the QDA model to perform better, more balanced independent variables need to be included in the model. An accurate classification is not possible from the information provided by the independent variables that were available. This can also be seen in the fact that the QDA model showed the lowest discrimination ability score with an AUC of 0.825.



2.3 Conclusion/Discussion

Future model expansion could occur from either additional variables or different algorithms. Intuitively, salary and cost of living vary significantly by location within the country and urban vs. rural. Having location information for these observations could improve performance of all models. Throughout this paper, numerous interpretable models were implemented, but less interpretable and potentially higher AUC models could be considered including neural networks and gradient boosted machines.

Due to the data collection methods being performed observationally, rather than experimentally, it should be stressed that all potential relationships identified are correlations not causations. It should also be noted that although the data came from the census, we do not know if the rows provided in the data set are a random sample. As a result, we cannot confidently apply our findings to the entire United States 1994 population. In summary, various models ranging in complexity and interpretability were produced to generate predictions whether this sample of citizens earned more or less than $50k a year.

Appendix

R Code

Code as well as additional analysis are available in our github repository:

https://github.com/rickfontenot/Predicting_Income

Data Description

Covariate	Type	Description	Nominal Levels
Age	Continuous	Age of an individual	
Workclass	Categorical	Socioecominc status	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay and Never-worked
FNLWGT	Continuous	Approximately how many people the Census believes this entry represents	
Education	Categorical	Highest level of education achieved	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th and Masters
Education Num	Continuous	Highest level of education achieved	1st-4th, 10th, Doctorate, 5th-6th and Preschool
Marital-status	Categorical	Marital status of an individual	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent and Married-AF-spouse
Occupation	Categorical	An individuals job or profession	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv and Armed-Forces.
Relationship	Categorical	An individuals connection to others	Wife, Own-child, Husband, Not-in-family, Other-relative and Unmarried
Race	Categorical	An individual's physical characteristics	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other and Black
Sex	Categorical	An individual's bilogical sex	Female and Male
Capital Gain	Continuous	Profit earned on the sale of an asset	
Capital Loss	Continuous	Profit loss on the sale of an asset	
Hours per Week	Continuous	An individual's hours worked per week	
Native Country	Categorical	An individual's country of origin	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

Exploratory Data Analysis Details

The UCI dataset includes “?” for fields with missing values. We replaced this with NA to do an assessment and found that 5.6% observations are missing values for both occupation & workclass. 1.7% are missing native_country information.

Through discussion we decided not to drop observations with NA for occupation and workclass because the fact that the values are missing could be useful data for making predictions. For instance, a non-response could mean that their occupation could be hard to nail down if the person works multiple low wage jobs simultaneously to make ends meet of if they are in a transition between occupations. This lack of response could be a useful predictor.

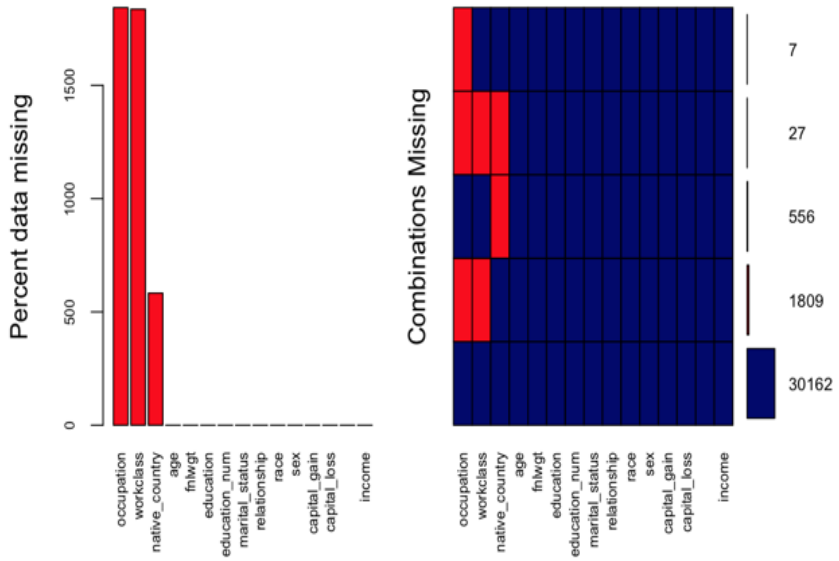


Figure EDA1: Dataset missing values.

Categorical Variable Summaries:
Upon reviewing distributions for the categorical variable levels, significant insights include:

- Work class is 70% Private sector, may want to explore grouping various government and self-employed categories
- Education has a factor level ordering issue, need to re-assign levels to proper order and study it's correlation to the numerical Education variable available
- Race is 85% white, 9.6% Black, does not seem to align with make-up of country. Is the census data skewed, or are responses inaccurate?
- Sex is 67% Male
- Native Country is 90% U.S. and has a lot of other level with small percentages, may want to reduce to other, or region

Characteristic	N = 32,561
workclass	
Federal-gov	960 (3.1%)
Local-gov	2,093 (6.8%)
Never-worked	7 (<0.1%)
Private	22,696 (74%)
Self-emp-inc	1,116 (3.6%)
Self-emp-not-inc	2,541 (8.3%)
State-gov	1,298 (4.2%)
Without-pay	14 (<0.1%)
Unknown	1,836
education	
10th	933 (2.9%)
11th	1,175 (3.6%)
12th	433 (1.3%)
1st-4th	168 (0.5%)
5th-6th	333 (1.0%)
7th-8th	646 (2.0%)
9th	514 (1.6%)
Assoc-acdm	1,067 (3.3%)
Assoc-voc	1,382 (4.2%)
Bachelors	5,355 (16%)
Doctorate	413 (1.3%)
HS-grad	10,501 (32%)
Masters	1,723 (5.3%)
Preschool	51 (0.2%)
Prof-school	576 (1.8%)
Some-college	7,291 (22%)
marital_status	
Divorced	4,443 (14%)
Married-AF-spouse	23 (<0.1%)
Married-civ-spouse	14,976 (46%)
Married-spouse-absent	418 (1.3%)
Never-married	10,683 (33%)
Separated	1,025 (3.1%)
Widowed	993 (3.0%)
occupation	
Adm-clerical	3,770 (12%)
Armed-Forces	9 (<0.1%)
Craft-repair	4,099 (13%)
Exec-managerial	4,066 (13%)
Farming-fishing	994 (3.2%)
Handlers-cleaners	1,370 (4.5%)
Machine-op-inspct	2,002 (6.5%)
Other-service	3,295 (11%)
Priv-house-serv	149 (0.5%)
Prof-specialty	4,140 (13%)
Protective-serv	649 (2.1%)
Sales	3,650 (12%)
Tech-support	928 (3.0%)
Transport-moving	1,597 (5.2%)
Unknown	1,843
relationship	
Husband	13,193 (41%)
Not-in-family	8,305 (26%)
Other-relative	981 (3.0%)
Own-child	5,068 (16%)
Unmarried	3,446 (11%)
Wife	1,568 (4.8%)
race	
Amer-Indian-Eskimo	311 (1.0%)
Asian-Pac-Islander	1,039 (3.2%)
Black	3,124 (9.6%)
Other	271 (0.8%)
White	27,816 (85%)
sex	
Female	10,771 (33%)
Male	21,790 (67%)
native_country	
Cambodia	19 (<0.1%)
Canada	121 (0.4%)
China	75 (0.2%)
Columbia	59 (0.2%)
Cuba	95 (0.3%)
Dominican-Republic	70 (0.2%)
Ecuador	28 (<0.1%)
El-Salvador	106 (0.3%)
England	90 (0.3%)
France	29 (<0.1%)
Germany	137 (0.4%)
Greece	29 (<0.1%)
Guatemala	64 (0.2%)
Haiti	44 (0.1%)
Holand-Netherlands	1 (<0.1%)
Honduras	13 (<0.1%)
Hong	20 (<0.1%)
Hungary	13 (<0.1%)
India	100 (0.3%)
Iran	43 (0.1%)
Ireland	24 (<0.1%)
Italy	73 (0.2%)
Jamaica	81 (0.3%)
Japan	62 (0.2%)
Laos	18 (<0.1%)
Mexico	643 (2.0%)
Nicaragua	34 (0.1%)
Outlying-US(Guam-USVI-etc)	14 (<0.1%)
Peru	31 (<0.1%)
Philippines	198 (0.6%)
Poland	60 (0.2%)
Portugal	37 (0.1%)
Puerto-Rico	114 (0.4%)
Scotland	12 (<0.1%)
South	80 (0.3%)
Taiwan	51 (0.2%)
Thailand	18 (<0.1%)
Trinadad&Tobago	19 (<0.1%)

Figure EDA2: Categorical variable summary.

Numerical variable summaries:

- Age appears to have a normal distribution but is truncated on the left since younger people are not employed and included in this set. A log transformation reduces the right skew some but does not overcome the truncated left tail. We will assume age is from a normal distribution as needed to meet assumptions of tests/models
- Fnlwgt is also skewed and could benefit from transformation. However the information represents weight of how much of state population may be similar to the demographics of each observation. The scale of this weighting varies by state and we do not have location

information to do any normalization on ranges. This information is not a predictor of that specific observations income so we will drop from modeling

- Capital gain & loss are almost all zero, then have long right skews. It may be helpful to evaluate converting these to categorical Yes/No variables
- Education correlates perfectly with the categorical variable and represents the ordering.

There are spikes at 9, 10, 13 which correspond to high school, some college, and bachelors degree.

- Hours per week spike has a large spike at 40 but is relatively normally distributed

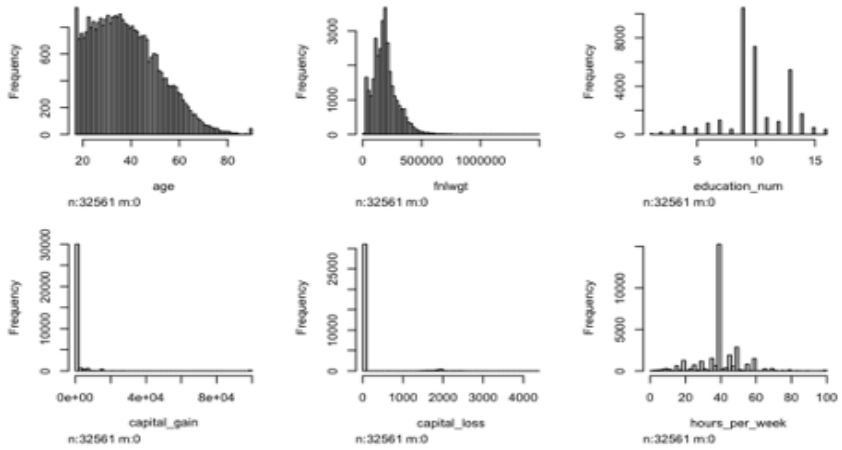
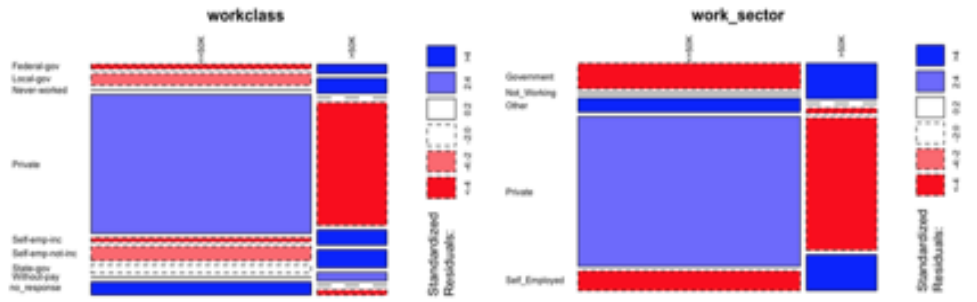


Figure EDA3: Numerical variable distributions.

Categorical Variable Relationships to Income:
For each categorical variable, we created mosaic plots vs. income to check for significant differences by level, explore level re-groupings and reductions. For variables with potential level adjustments, we compared Chi-Square tests before and after as well as the effects on simple logistic regression performance with just the one variable change. In cases where Chi-square improved or was similar and regression performance improved or was similar, we chose to use the reduced level factor for model performance and simpler interpretation.

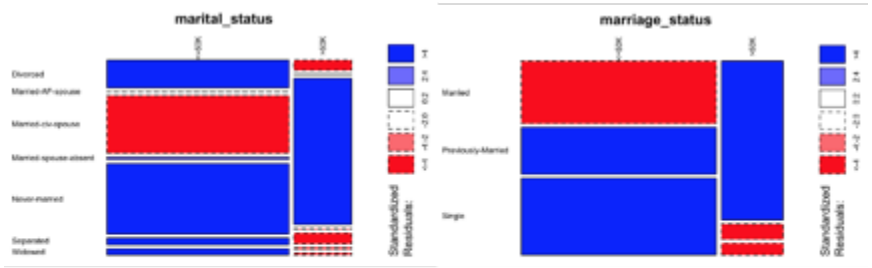
Workclass:
Almost all levels show a significant difference with income and appear to be in same direction for broad groups of sectors. Upon grouping to Private, Government, Self Employed, Not Working, and Other into a new variable “work_sector” the X-square slightly decreased from 828 to 565 (p-value still 2.2e-16) but model performance improved.




```
AIC: 16511 vs original 16561 -->improved
Accuracy : 0.8201 vs original 0.8188 -->improved
Sensitivity : 0.9395 vs original 0.9369 --> improved
Specificity : 0.4571 vs original 0.4584 -->dropped just a little
```

Figure EDA4: Workclass relationship to income.

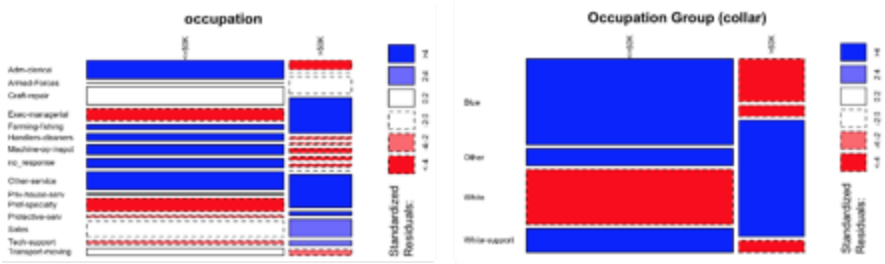
Marital Status:
All levels show a significant difference with income and appear to be in same direction for broad groups of status. Upon grouping to Married, Previously Married, and Single into a new variable "marriage_status" the X-square did not significantly change from 6518 to 6509 (p-value still 2.2e-16) and model performance improved.




```
AIC: 14470 vs original 16561 -->improved
Accuracy : 0.832 vs original 0.8188 -->improved
Sensitivity : 0.9237 vs original 0.9369 --> dropped a little
Specificity : 0.5656 vs original 0.4584 -->improved
```

Figure EDA5: Marital Status relationship to income.

Occupation:
Most levels show a significant difference with income and appear to be in same direction for groups which are office work (white-collar), physical labor intensive (blue collar) or other (primarily the no_response observations). Within the white-collar office jobs, Admin-clerical had a different relationship to income than managerial and technical occupations, so it was set into its own level for support role within white collar jobs. Upon grouping into a new variable "collar" the X-square dropped some from 3745 to 2884 (p-value still 2.2e-16) but model performance improved.




```
AIC: 16362 vs original 16561 -->improved
Accuracy : 0.8248 vs original 0.8188 -->improved
Sensitivity : 0.9385 vs original 0.9369 --> improved
Specificity : 0.4724 vs original 0.4584 -->improved
```

Figure EDA6: Occupation relationship to income.

Education:
After re-ordering levels in education, it is perfectly correlated to the numerical version, education_num. See subsequent section for exploration of the numerical version vs. income.

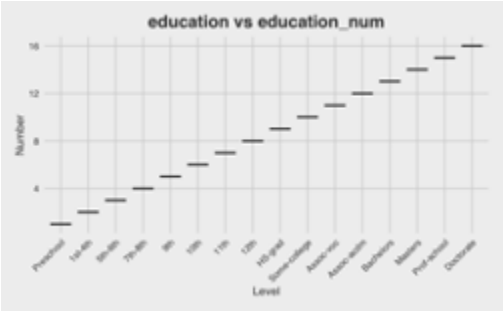


Figure EDA7: Education relationship to education_num.

Relationship:
All levels are significant. It's possible the information in this variable may be redundant with the combination of marriage status and sex. No level reduction was performed but we will evaluate significance with combination of these three variables.

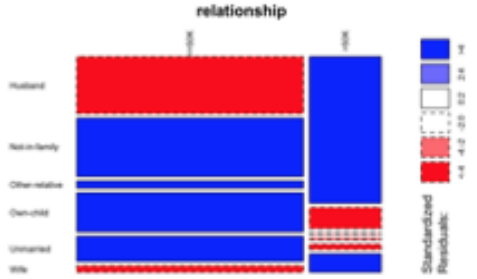


Figure EDA8: Relationship status relationship to income.

Numerical Variable Relationships to Income:

The proportion of people with income above $50K was calculated and plotted vs each numerical variable to examine relationships and search for any non-linear trends. In cases where the relationship appears non-linear, regression models were built adding one quadratic term at a time to look for significance and improvements in fit. In cases where second or third order terms improved the fit, these terms were added to the dataset to include in model evaluations.

Age:
Age exhibits a non-linear relationship to proportion above $50K, through iterative regression models both the second and third order terms improved the fit, a fourth order term was not significant.

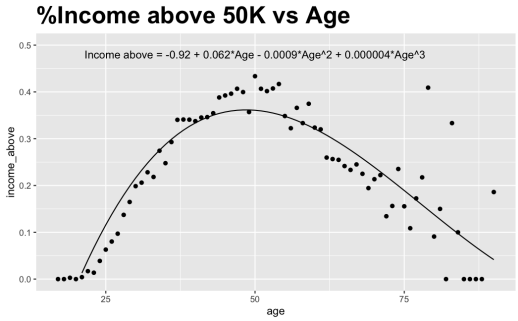


Figure EDA9: Age relationship to income

Education:
Education exhibits a non-linear relationship to proportion above $50K, through iterative regression models the second term improved the fit, a third order term was not significant.

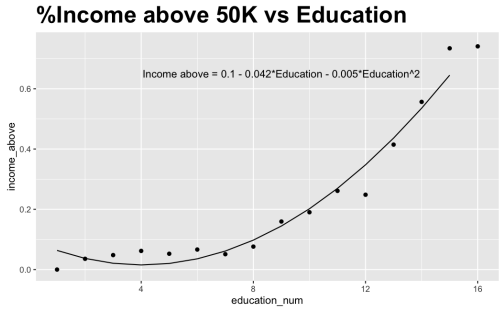


Figure EDA10: Education relationship to income

Hours per Week:
Hours per week exhibits a non-linear relationship to proportion above $50K, through iterative regression models the second term improved the fit, a third order term was not significant. Note that this fit is not as significant as fits obtained on other variables. Hours per week may perform better in non-parametric modeling.

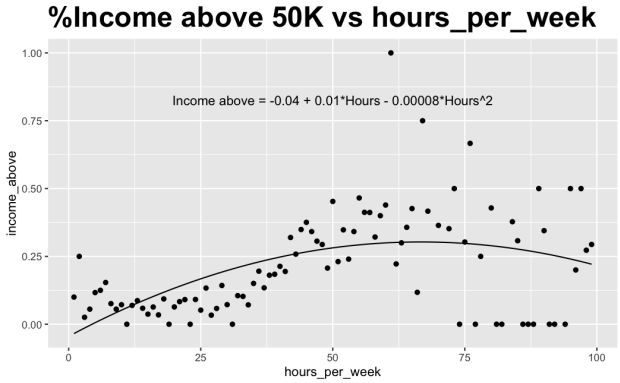


Figure EDA11: Hours per week relationship to income.

Capital Gain and Loss:
Capital gain and loss were binned since the distribution is primarily zero's with a small proportion in long right skews with many different values. They do not exhibit a linear relationship but no regression fits or transformations were immediate to improve linear relationships for modeling.

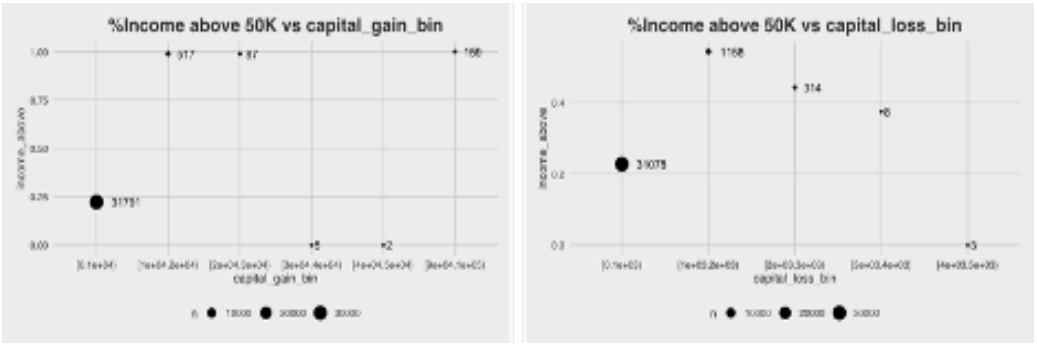


Figure EDA12: Capital gain and capital loss relationship to income.

Simplified Logistic Regression

```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -6.889e+00  2.822e-01 -24.413  < 2e-16 ***
age                              2.195e-02  1.670e-03  13.147  < 2e-16 ***
education_num                    2.905e-01  9.407e-03  30.883  < 2e-16 ***
raceAsian-Pac-Islander           3.692e-01  2.653e-01   1.392 0.163902
raceBlack                        4.732e-01  2.541e-01   1.862 0.062573 .
raceOther                        2.204e-02  3.728e-01   0.059 0.952853
raceWhite                        5.464e-01  2.434e-01   2.245 0.024753 *
sexMale                          1.722e-01  5.563e-02   3.095 0.001965 **
capital_gain                     3.188e-04  1.110e-05  28.724  < 2e-16 ***
capital_loss                     6.841e-04  4.124e-05  16.590  < 2e-16 ***
hours_per_week                   3.006e-02  1.698e-03  17.701  < 2e-16 ***
marriage_statusPreviously-Married -2.139e+00  6.443e-02 -33.205  < 2e-16 ***
marriage_statusSingle            -2.758e+00  7.016e-02 -39.316  < 2e-16 ***
collarOther                      2.716e-01  8.080e-02   3.362 0.000775 ***
collarWhite                      8.125e-01  4.796e-02  16.941  < 2e-16 ***
collarWhite-support              4.125e-01  7.652e-02   5.390 7.03e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure SLR1: Summary of estimates and significance for coefficients of selected simplified logistic regression model.

```
          age    education_num             race              sex     capital_gain     capital_loss
     1.130835         1.273159         1.006939         1.342343         1.023477         1.008587
hours_per_week   marriage_status           collar
     1.083082         1.182923         1.116378
```

Figure SLR2: VIF of Selected simplified model predictors.

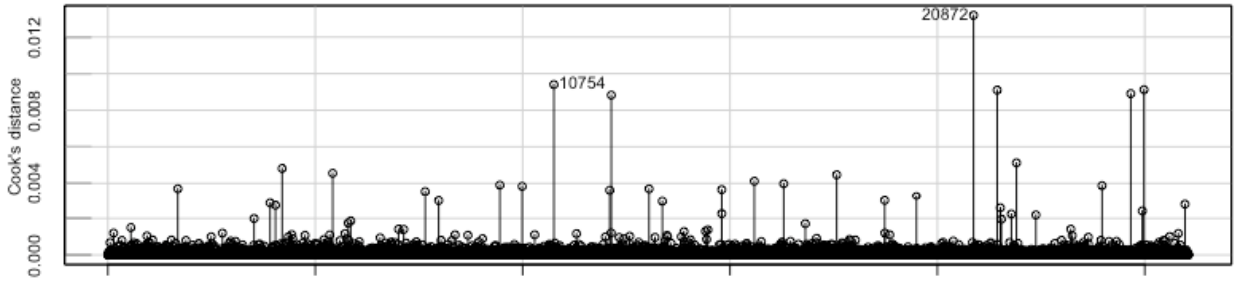


Figure SLR3: Cook's D for each observation the model was trained on.

	Model 1	Model 2
(Intercept)	-6.90	-6.89
age	0.022	0.022
education_num	0.291	0.291
raceAsian-Pac-Islander	0.372	0.369
raceBlack	0.476	0.473
raceOther	0.0261	0.0220
raceWhite	0.550	0.546
sexMale	0.172	0.172
capital_gain	0.000324	0.000319
capital_loss	0.000685	0.000684
hours_per_week	0.0301	0.0301
marriage_statusPreviously-Married	-2.14	-2.14
marriage_statusSingle	-2.76	-2.76
collarOther	0.270	0.272
collarWhite	0.811	0.812
collarWhite-support	0.412	0.412

Figure SLR4: Comparison of Model1 which is the original model without the most influential point (observation 20872) to Model2 which is the original selected model with all observations.

Additional Influential Point Analysis Details:
To investigate the magnitude of the residuals, we can look at studentized residuals. To investigate the magnitude of the leverage, we can look at the hat-values. The metric used to consider both residual and leverage is Cook's D. Figure SLR5 visualizes magnitude of residual, leverage, and influence all at once by showing the Studentized Residuals vs. Hat-Values with circle size representing Cook's D. The comparative values of Studentized Residual, Hat-Value, and Cook's D for the five most influential observations can be found in Figure SLR6.

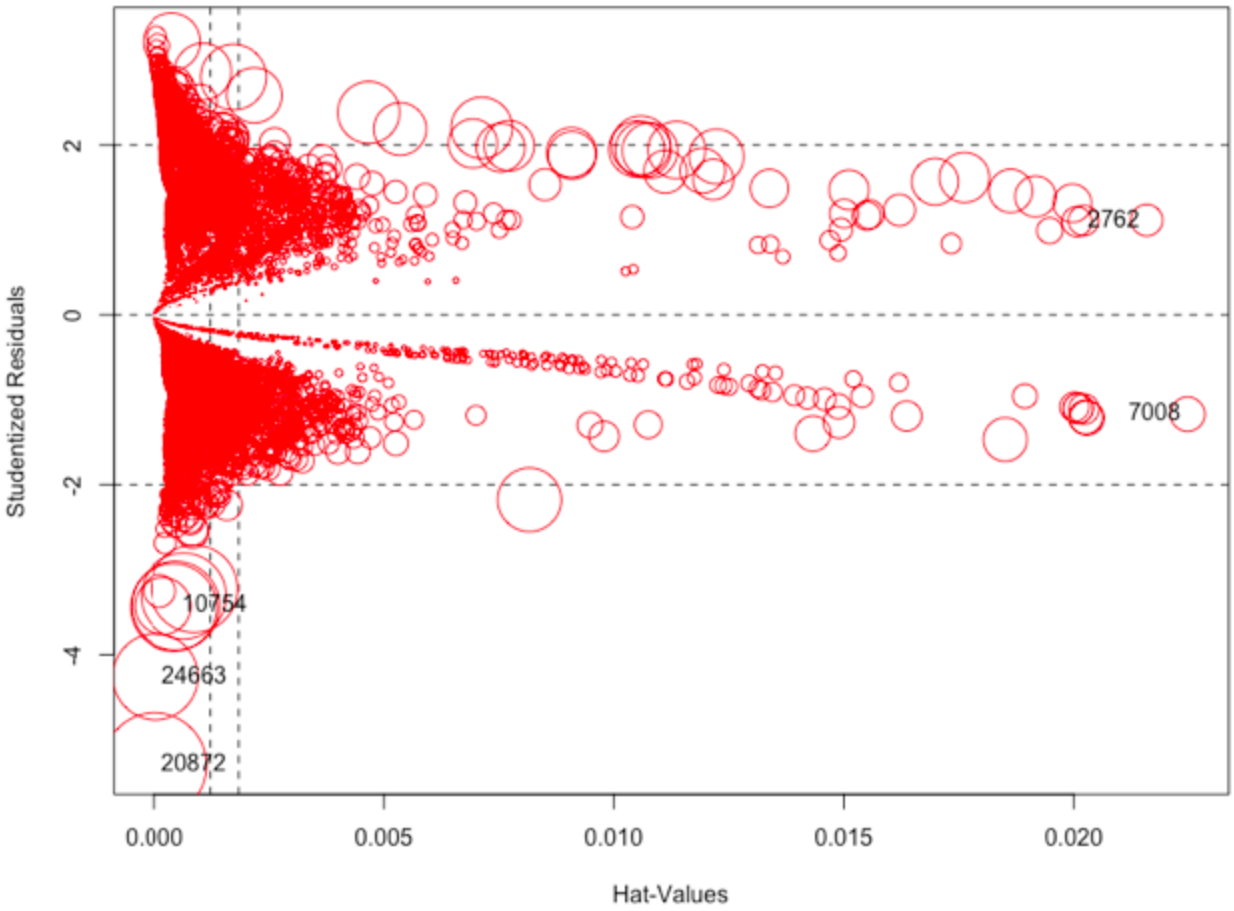


Figure SLR5: Studentized Residuals vs. Hat-Values (size of circle represents Cook's D).

	StudRes	Hat	CookD
2762	1.114890	2.158490e-02	0.001190015
7008	-1.168989	2.247268e-02	0.001409066
10754	-3.424315	4.621407e-04	0.009404128
20872	-5.302623	1.838911e-07	0.013192015
24663	-4.259101	1.761029e-05	0.008906418

Figure SLR6: Top 5 most influential points.

We can see from the figure that observations 2762 and 7008 have high leverage, while observations 10754, 24663, and 20872 have high residuals. None of these highlighted influential points were significant in both residual and leverage, according to Figure SLR5.

	Odds ratio	2.5 %	97.5 %
(Intercept)	0.001758885	0.001303336	0.00237366
age	1.022197733	1.018857535	1.02554888
education_num	1.337100559	1.312675062	1.36198055
raceAmer-Indian-Eskimo	0.579032421	0.359381650	0.93293173
raceAsian-Pac-Islander	0.837654217	0.676391066	1.03736525
raceBlack	0.929412496	0.795415111	1.08598338
raceOther	0.591936171	0.339219985	1.03292390
sexMale	1.187906693	1.065200326	1.32474829
capital_gain	1.000318888	1.000297125	1.00034065
capital_loss	1.000684360	1.000603482	1.00076524
hours_per_week	1.030514199	1.027090147	1.03394966
marriage_statusPreviously-Married	0.117743146	0.103775782	0.13359040
marriage_statusSingle	0.063401614	0.055256485	0.07274738
collarOther	1.312062422	1.119902903	1.53719380
collarWhite	2.253456514	2.051288115	2.47554999
collarWhite-support	1.510582732	1.300188685	1.75502234

Figure SLR7: Odds Ratio and 95% confidence interval for each coefficient.

Interpretations of simplified logistic regression model:

We will assume the interpretations below are valid only when all other predictors are held constant:

Categorical Variable Interpretations:

- Race
    - The odds of earning over $50K for the Amer-Indian-Eskimo population is expected to be 42.09% less than the White population
    - The odds of earning over $50K for the Asian-Pac-Islander population is expected to be 16.24% less than the White population
    - The odds of earning over $50K for the Black population is expected to be 7.06% less than the White population
    - The odds of earning over $50K for the Other Race population is expected to be 40.81% less than the White population
- Sex
    - The odds of earning over $50K for the Male population is expected to be 18.79% higher than the Female population
- Marriage_status
    - The odds of earning over $50K for the Previously-Married population is expected to be 88.23% less than the Married population
    - The odds of earning over $50K for the Single population is expected to be 93.66% less than the Married population
- Collar

- The odds of earning over $50K for the Other-Collar population is expected to be 31.21% higher than the Blue-Collar population
- The odds of earning over $50K for the White-Collar population is expected to be 125.35% higher than the Blue-Collar population
- The odds of earning over $50K for the White-Support Collar population is expected to be 51.06% higher than the Blue-Collar population

Continuous Variable Interpretations:

- Age
  - For every 1 year increase in age, we expect the odds of earning over $50K to increase by 2.22%
- Education_num
  - For every 1 year increase in education, we expect the odds of earning over $50K to increase by 33.71%
- Capital_Gain
  - For every $1K increase in Capital_Gain, we expect the odds of earning over $50K to increase by 37.55%
- Capital_Loss
  - For every $1K increase in Capital_Loss, we expect the odds of earning over $50K to increase by 98.20%
- Hours_per_week
  - For every 1 hour increase in hours worked per week, we expect the odds of earning over $50K to increase by 3.05%

Figure SLR8: Interpretation of coefficients.

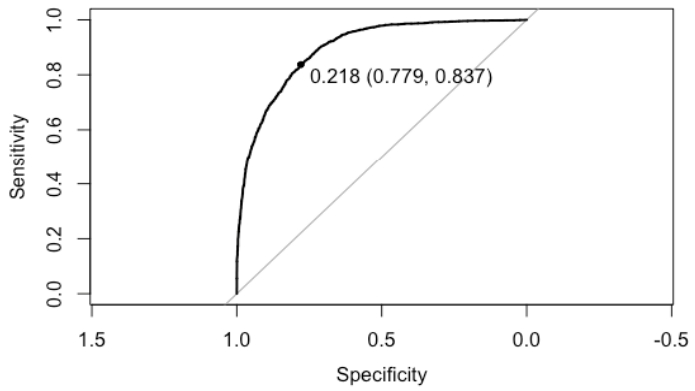


Figure SLR9: ROC Curve for simplified logistic regression model.

```
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K  3835  260
     >50K   1086 1331

               Accuracy : 0.7933
                 95% CI : (0.7833, 0.8031)
    No Information Rate : 0.7557
    P-Value [Acc > NIR] : 3.444e-13

                  Kappa : 0.5239

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7793
            Specificity : 0.8366
         Pos Pred Value : 0.9365
         Neg Pred Value : 0.5507
             Prevalence : 0.7557
         Detection Rate : 0.5889
   Detection Prevalence : 0.6288
      Balanced Accuracy : 0.8079

       'Positive' Class : <=50K
```

Figure SLR10: Confusion matrix summary for simplified logistic regression model.

Complex Logistic Regression

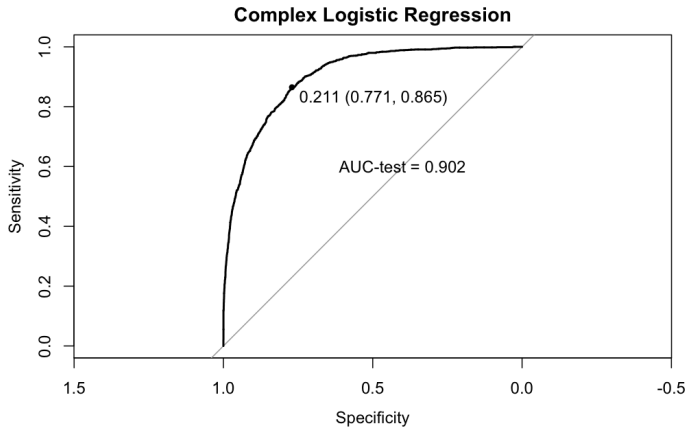


Figure CLR1: ROC Curve for complex logistic regression model.

```
          Reference
Prediction <=50K >50K
     <=50K  3792  215
     >50K   1129 1376

               Accuracy : 0.7936
                 95% CI : (0.7836, 0.8034)
    No Information Rate : 0.7557
    P-Value [Acc > NIR] : 2.217e-13

                  Kappa : 0.532

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7706
            Specificity : 0.8649
         Pos Pred Value : 0.9463
         Neg Pred Value : 0.5493
             Prevalence : 0.7557
         Detection Rate : 0.5823
   Detection Prevalence : 0.6153
      Balanced Accuracy : 0.8177

       'Positive' Class : <=50K
```

Figure CLR2: Confusion matrix summary for complex logistic regression model.

Quadratic Discriminant Analysis

```
        Generalized Shapiro-Wilk test for Multivariate
Normality by Villasenor-Alva and Gonzalez-Estrada

data:  multivariate_sample
MVW = 0.68754, p-value < 2.2e-16
```

Figure QDA1: Shapiro-Wilk test results.

```
      Box's M-test for Homogeneity of Covariance Matrices

data:  boxm_test[, c(1:5)]
Chi-Sq (approx.) = 76481, df = 15, p-value < 2.2e-16
```

Figure QDA2: Box's M-test results.

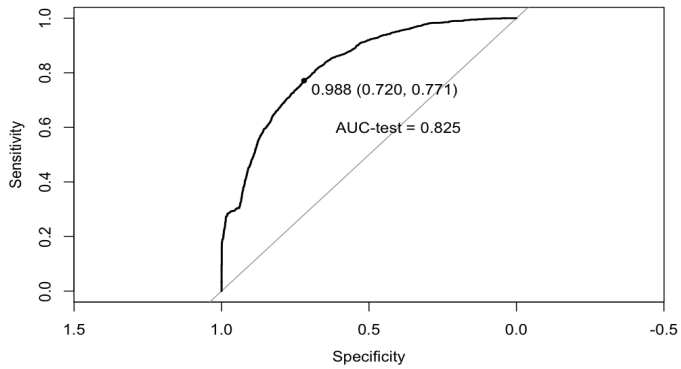


Figure QDA4: ROC curve for quadratic discriminant analysis model.

```
          Reference
Prediction <=50K >50K
     <=50K  3452  340
     >50K   1469 1251

               Accuracy : 0.7222
                 95% CI : (0.7112, 0.7331)
    No Information Rate : 0.7557
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3933

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7015
            Specificity : 0.7863
         Pos Pred Value : 0.9103
         Neg Pred Value : 0.4599
             Prevalence : 0.7557
         Detection Rate : 0.5301
   Detection Prevalence : 0.5823
      Balanced Accuracy : 0.7439

       'Positive' Class : <=50K
```

Figure QDA3: Confusion matrix summary for quadratic discriminant analysis model.

Random Forest

Random forests can handle any type of numeric variables, but categorical variables must be encoded numerically in some form. One approach that was implemented is caret's default version of one-hot encoding. In this method, each level of a category is represented by a new variable. If the category is associated with the new row, it receives a 1, otherwise a 0. Target encoding, or mean encoding, was another technique that was tested. Target encoding takes each level of a given variable and represents it as a number based off its relationship with the response variable. The stronger relationship the level has with the variable, the higher the number provided. To prevent data leakage, variables were only target encoded based on the training set. Then, once the relationship was determined and classes represented by a number, the classes in the test set were assigned their matching values. The advantage of this method is that target encoding does not increase the dimensionality of the dataset, unlike one hot encoding. This was particularly beneficial for the "native country" variable, which had many levels.

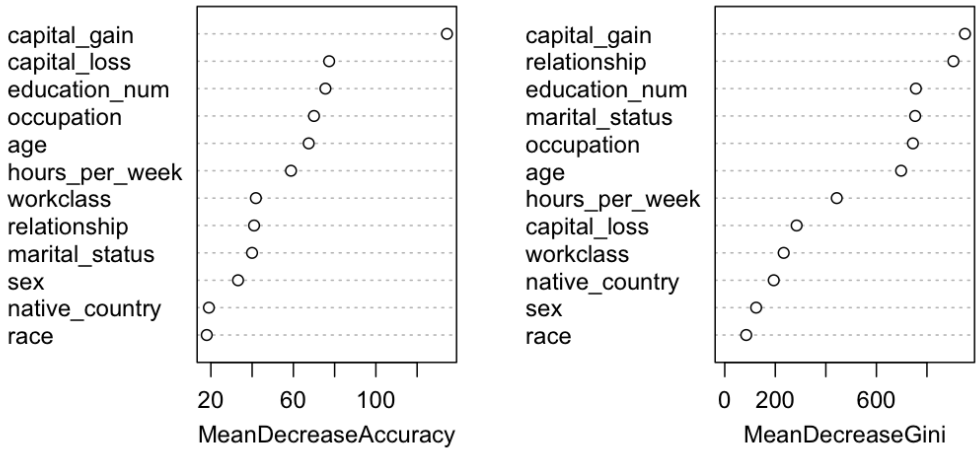


Figure RF1: Variable importance plot

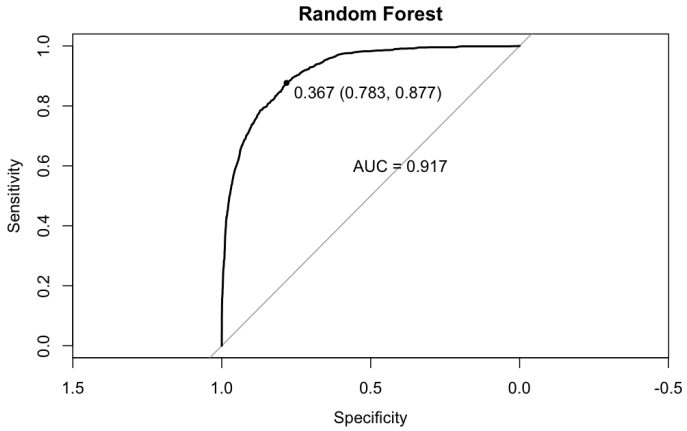


Figure RF2: ROC Curve for random forest model

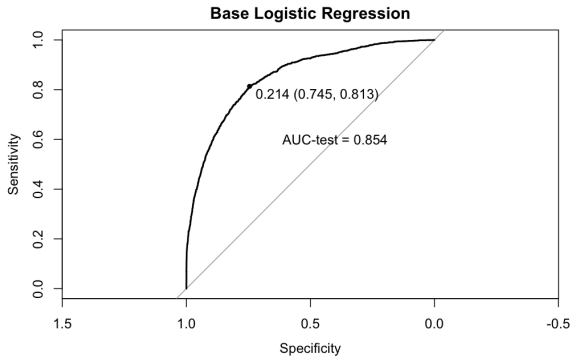
```
          Reference
Prediction <=50K >50K
     <=50K  3851  196
     >50K   1070 1395

               Accuracy : 0.8056
                 95% CI : (0.7958, 0.8151)
    No Information Rate : 0.7557
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.556

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7826
            Specificity : 0.8768
         Pos Pred Value : 0.9516
         Neg Pred Value : 0.5659
             Prevalence : 0.7557
         Detection Rate : 0.5914
   Detection Prevalence : 0.6215
      Balanced Accuracy : 0.8297

       'Positive' Class : <=50K
```

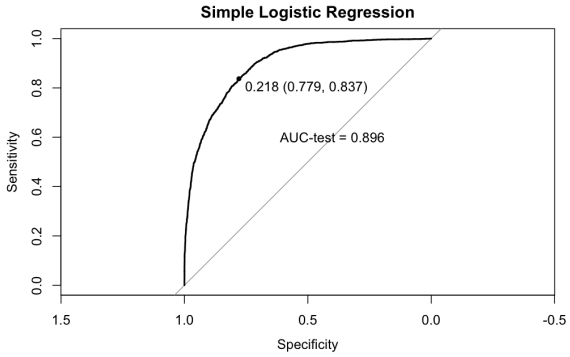
Figure RF3: Confusion matrix summary for random forest model

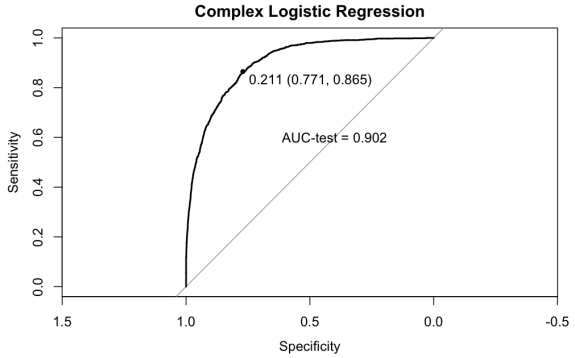
Final Model Comparison - Additional information

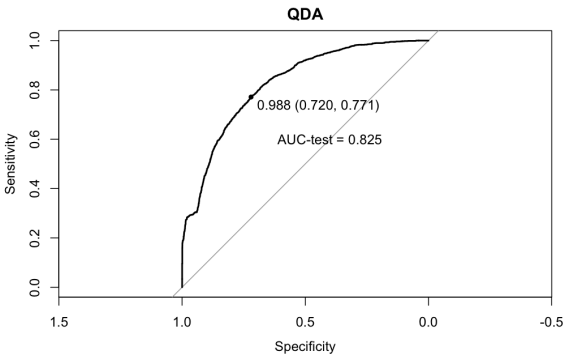
		Base Logistic Regression	Simple Logistic Regression	Complex Logistic Regression	QDA	Random Forest
	AUC (test set)	**0.854**	**0.896**	**0.902**	**0.825**	**0.917**
	Accuracy	0.761	0.794	0.794	0.722	0.806
	Sensitivity	0.744	0.780	0.771	0.702	0.7826
	Specificity	0.813	0.834	0.865	0.786	0.8768
	Balanced Accuracy	0.779	0.807	0.818	0.744	0.8297
Original Variables	age	X	X	X	S	X
	workclass	X				X
	fnlwgt	X				
	education	X				
	education_num	X	X	X	S	X
	marital_status	X				X
	occupation	X				X
	relationship	X				X
	race	X	X	X		X
	sex	X	X	X		X
	capital_gain	X	X	X	S	X
	capital_loss	X	X	X	S	X
	hours_per_week	X	X	X	S	X
	native_country	X				X
new features	work_sector					
	marriage_status		X	X		
	collar		X	X		
	age^2			X		
	age^3			X		
	education_num^2			X		
	hours_per_week^2			X		

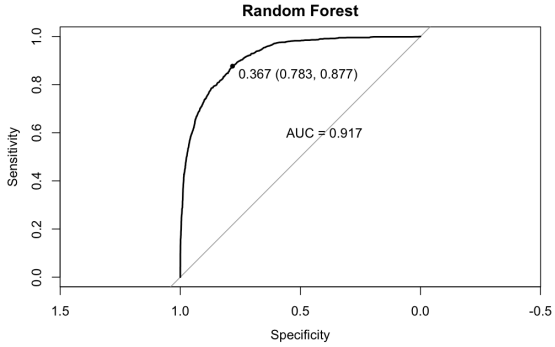
X = not significant S = scaled











References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.