**Who earned more than 50K in 1994?**

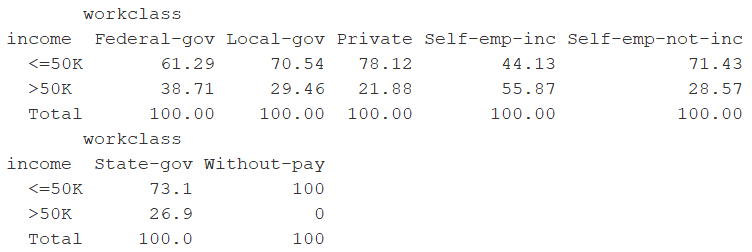
**Question of Interest**

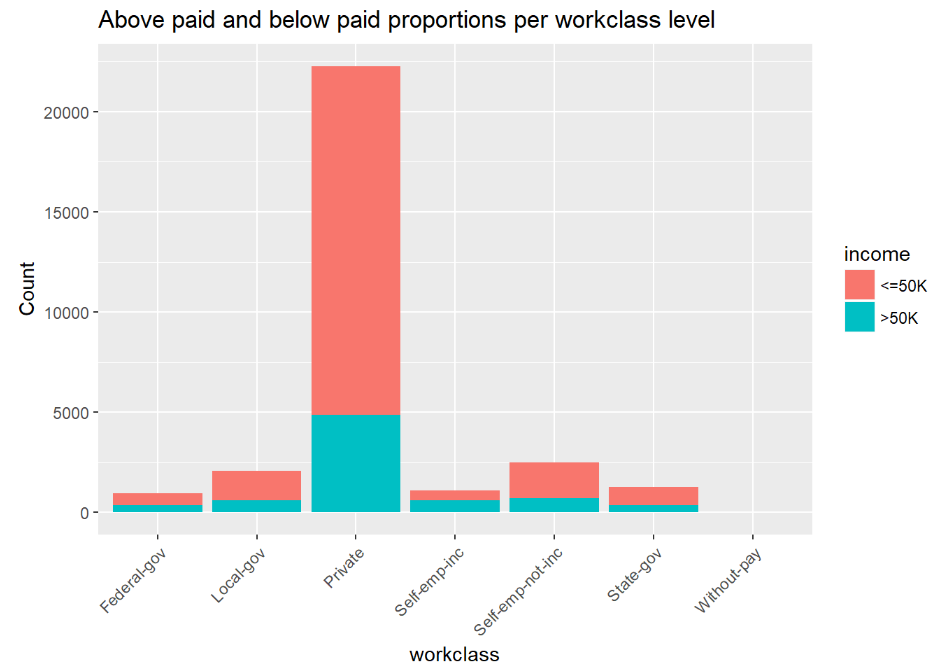
What factors or their weighted combinations can predict whether a person made over $50K a year in 1994 based on the given data.

**Exploratory Data Analysis (EDA)**

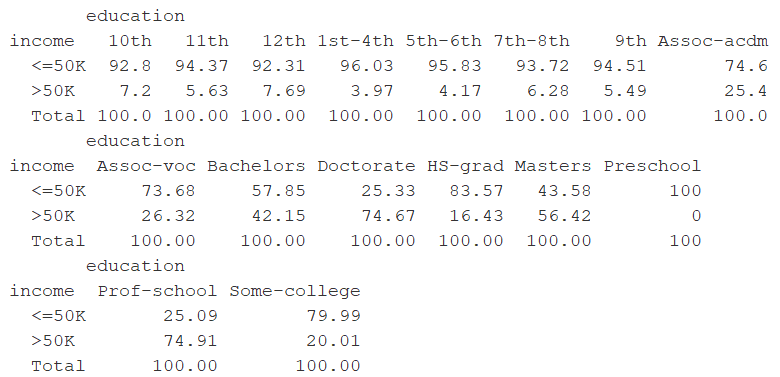
To better understand the factors that were most influential in resultant income of greater than $50K in 1994, we first explored the data to investigate the distributions of individual variables with respect to income category and in total. We further analyzed the relationship of those individual variables with the income category and the correlation between the variables themselves. The training and test sets are cleaned up first and the six new variables were created. Details of the EDA to assess the impact of explanatory variables on the likelihood of the high income are briefly summarized below.

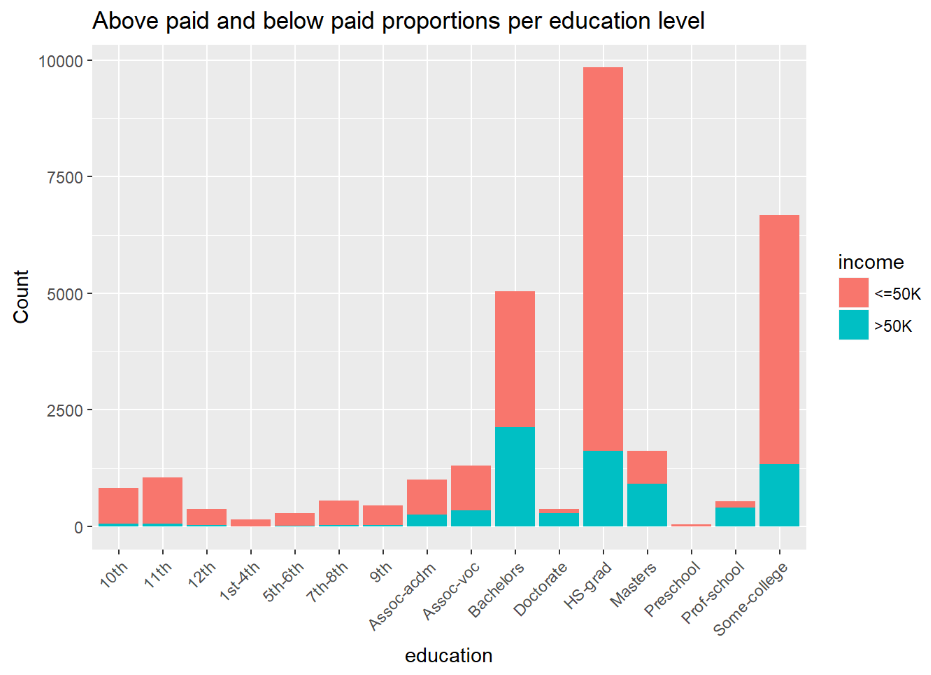
**Workclass:** The dataset comprises of 74% of the people who worked in Private sector. Incorporated self-employed workers had highest percentage (more than half, 56%) that earned more than 50K.



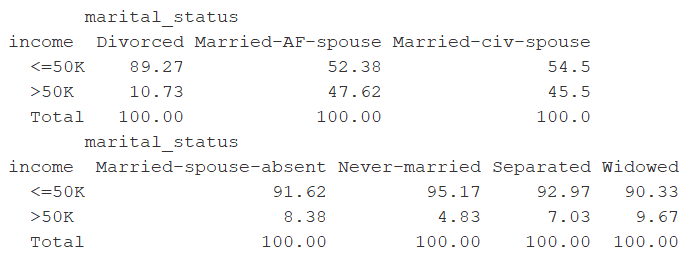


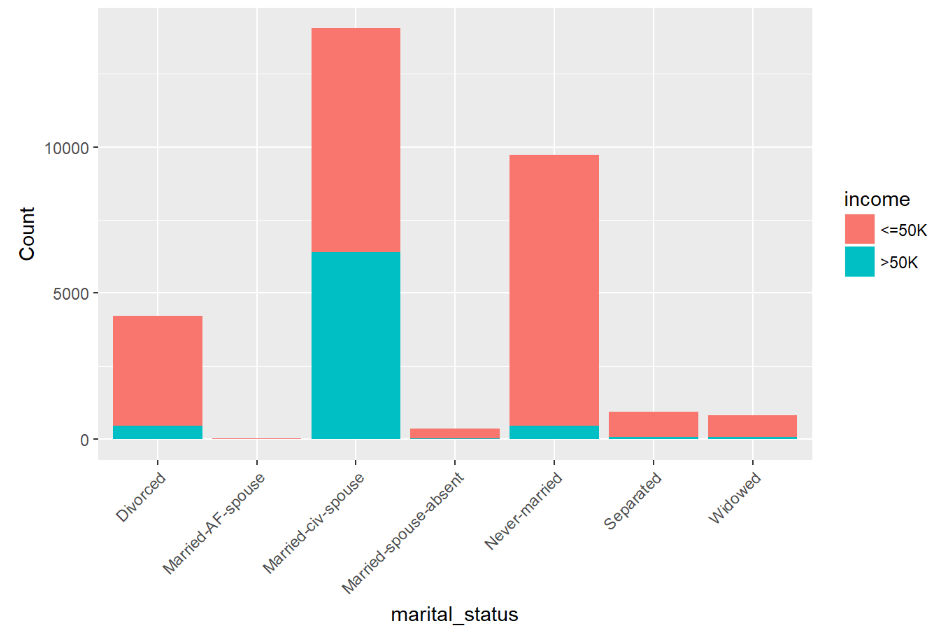
#### Education: Data is dominated by HS-grad (). As education level increases, percentage of people earning >50K increases. Higher education degree levels like Bachelor to PhD significantly contribute to higher pay grades.



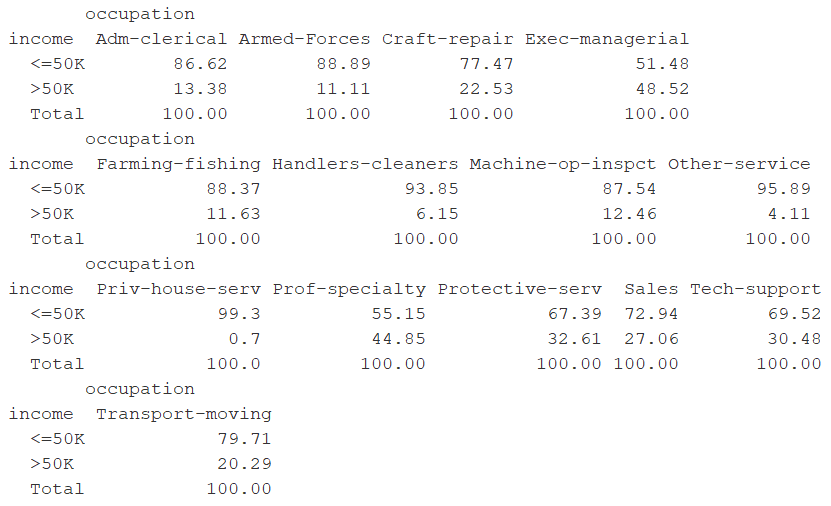


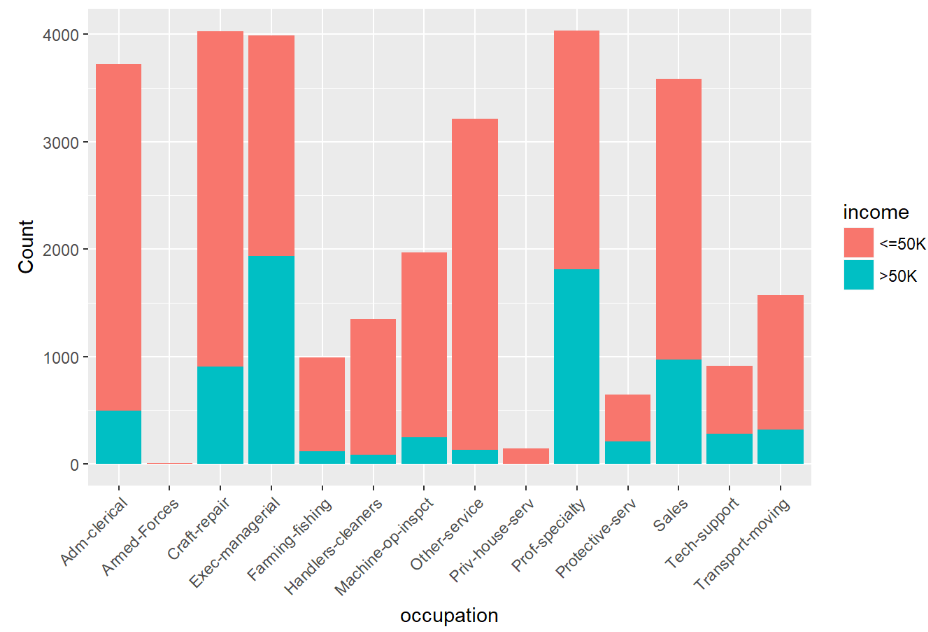
### **Marital status:** High percentage of >50K income earners are married civilians and living with spouse with good marital relationship.



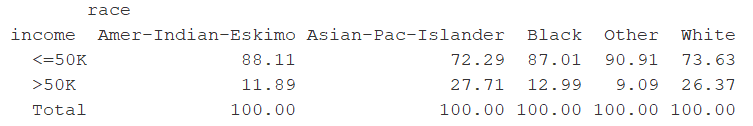


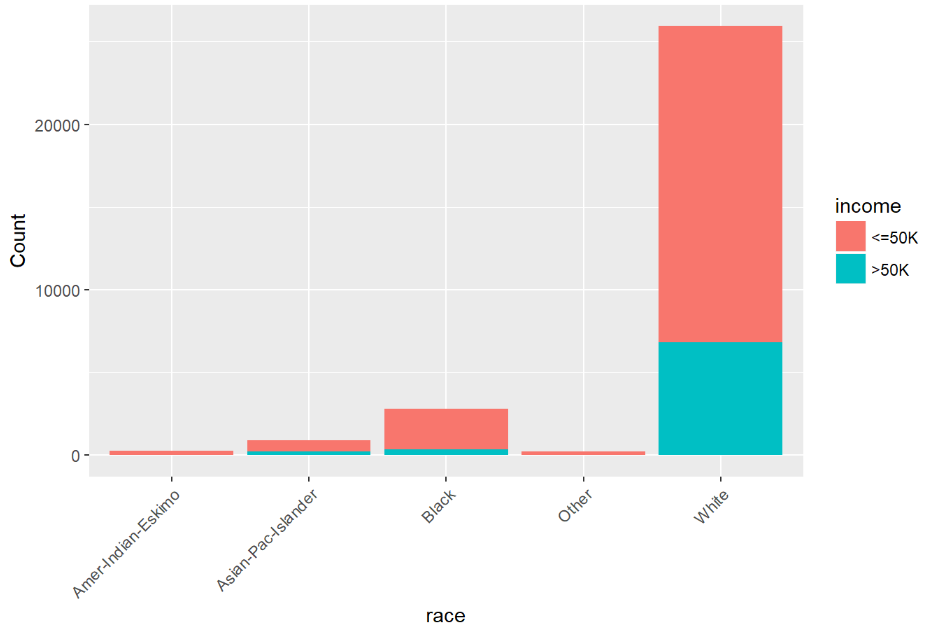
### **Occupation:** High percentage of >50K earner worked in executive managerial and professional specialty occupations**.**



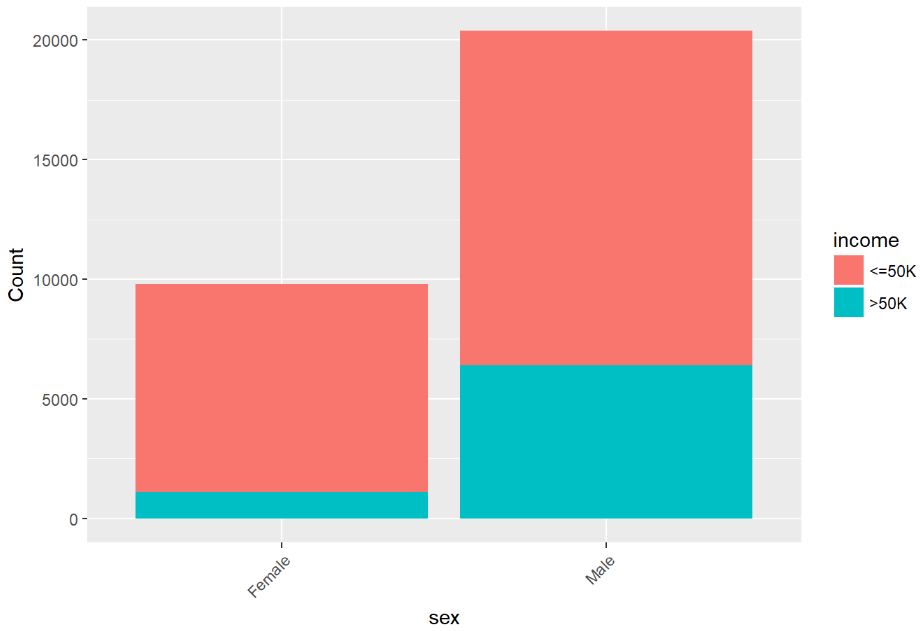
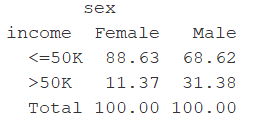


### **Race:** **Although** high percentage of high earners were Whites percentage of high earners was similar for Whites and Asian-Pac-Islanders.

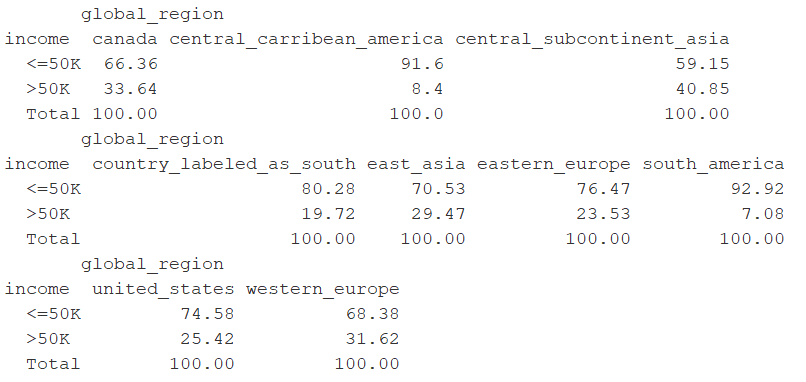


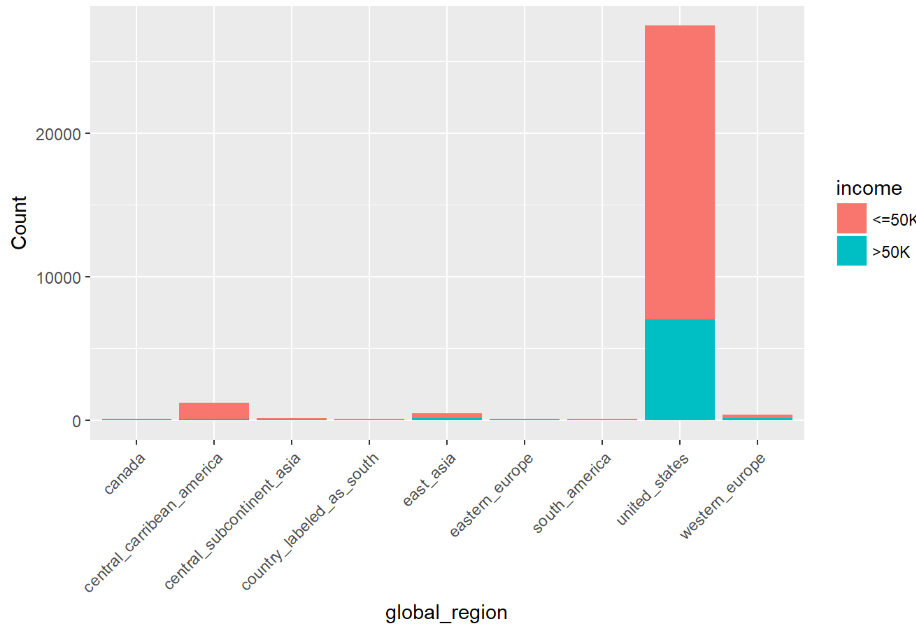


#### Sex: High paid workers were mostly Male.



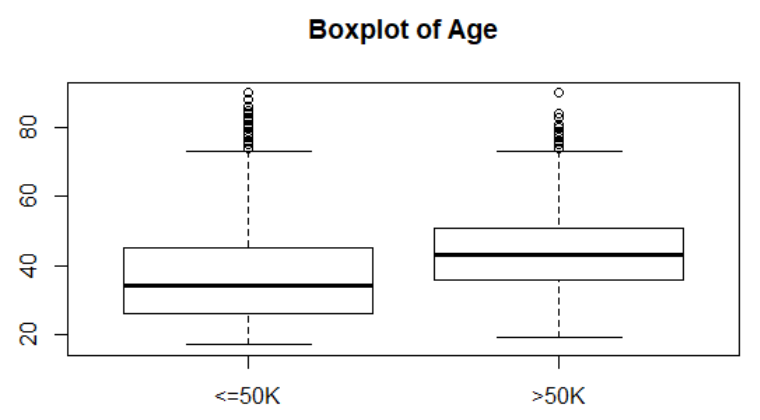
#### Global\_region: Although high percentage of high earners were from US percentage of high earners was higher for people from central\_subcontinent\_asia region.



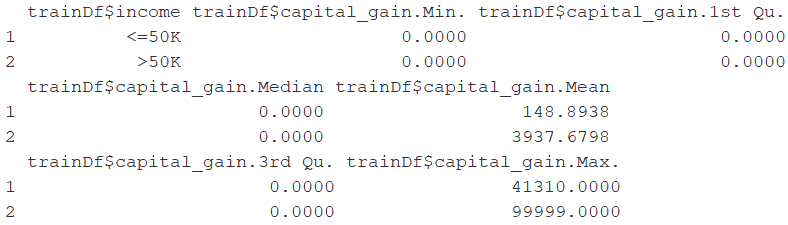


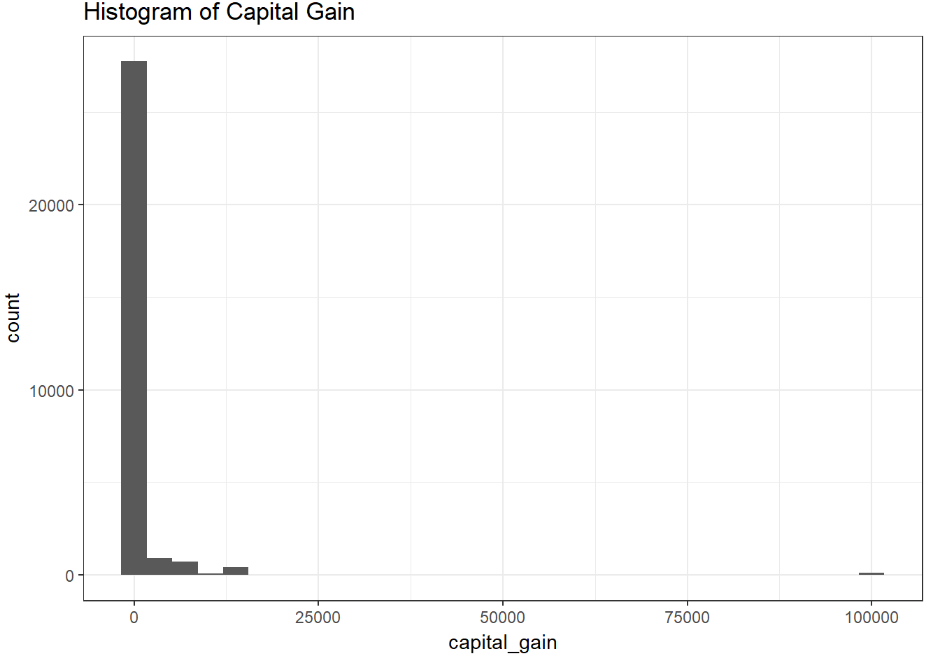
#### Summary stats for Continuous variables:

**Age:** Boxplot of Age shows people who earned >50k were between age of 36 to 51 with 43 as the median. Obviously, people between 17 to 24 were not earning high.

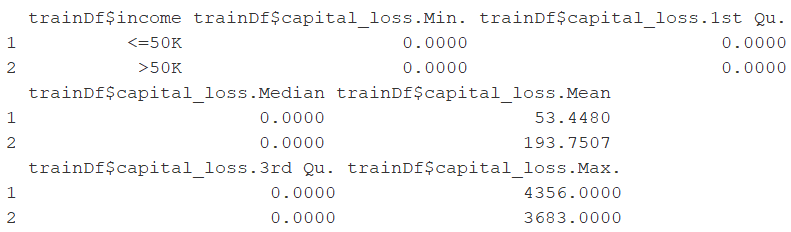


**Capital Gain:**





**Capital loss:**



**Predictive Model**

This dataset consists of majority of qualitative variables. Some of the qualitative variables are highly skewed. e.g. 86% of the data coming from “White” in “race” variable and 74 % of the data from “Private” sector “work class” variable.

We have different scales for different variables. To avoid few variables overweighing others in Principal Component Analysis (PCA) simply because of absolute scale we normalized all the variables. One of the limitation of PCA in this dataset is that the dependent variable levels are heavily skewed e.g. 75.1% observations from low income level and rest 24.9% from high income levels. So, we chose to use logistic regression to create a predictive model.

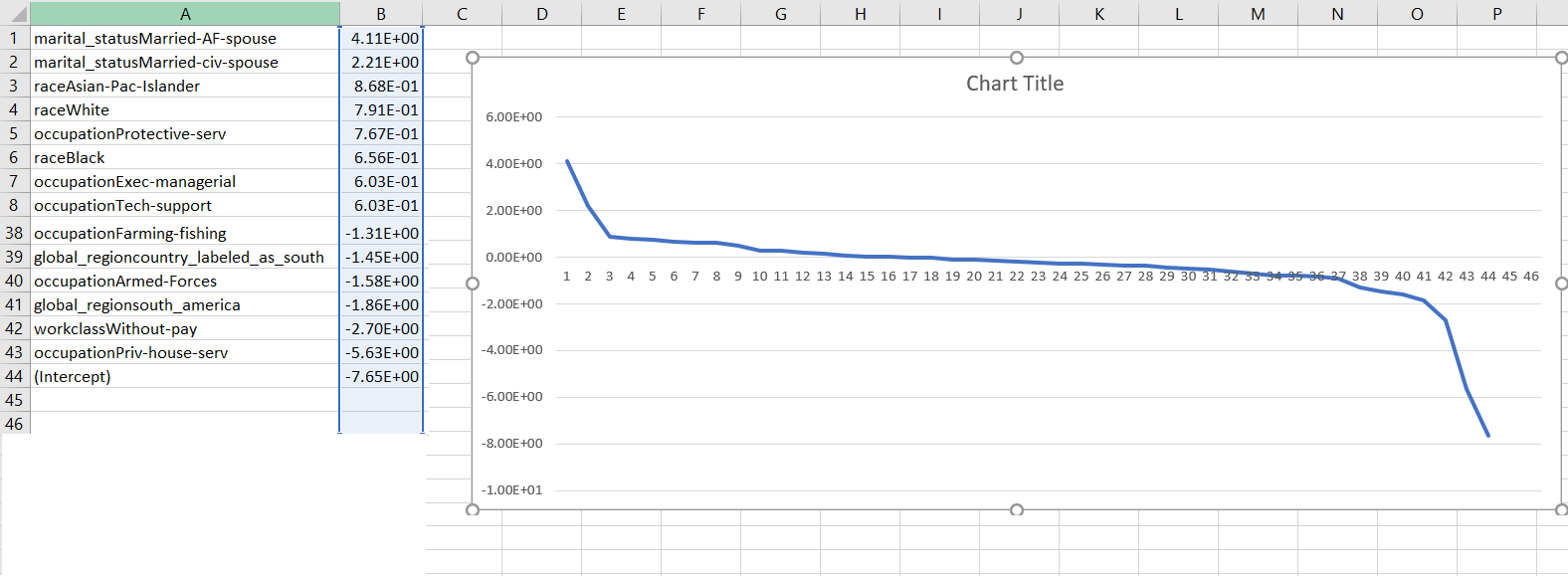
**Logistic regression:**

Since the response variable is binary, we are going to use logistic regression to predict who can earn >50k on test data set provided with this Adult census data. Based on EDA, we chose age, workclass, education\_num, marital\_status, occupation, race, sex, capital\_gain, capital\_loss, hours\_per\_week, global\_region as predictive explanatory variables in our logistic regression model. All the variables used in the model are statistically significant. Briefly writing the logistic regression equation below:

Ln(p/1-p) = y = -7.648 + (0.0341\*age) + (0.283\*education\_num) + ….

Both age (indirectly work experience) and education level increase the log odds of earning higher income in 1994.

The following image shows the coefficient values of the predictor variables and their plot in significance order which are most important predictors of earning higher income.



### **Conclusion:** In short, higher income community comprised of people who are married civilians and living with spouse with good marital relationship (“marital\_status” catergory), people of “race” Asian-Pac-Islander or White and people who are working (“occupation” category) in Protective services or Exec-managerial or Tech-support. Lower income community comprised of people who are without pay (“workclass”), people who are in the “occupation” of Priv-house-serv, people from south\_america “region”.