**Summary**

Utilizing the census data provided consisting of 13 demographic variables, a model was created in order to predict whether an individual makes over $50,000 a year. Both Random Forest and Logistic Regression models were tested with the Logistic Regression model proving to be the most accurate, with 84.8% accuracy.

**Exploration**

In trying to explore and understand the data, visualizations were created such as the image below. In this specific visualization, the blue represents the percentage of the population of each race that made over $50,000/year and the red represents the percentage that made under $50,000/year. This was very interesting in that it gave more information than was immediately available. When looking at just the total amount of people that make over $50,000/year, the naïve assumption would be that those that were identified as “White” more often make over $50,000/year. In looking at this visualization, it was shown that although there were more observations that were “White”, the “Asian-Pacific-Islander” had a higher percentage of people who made over $50,000.

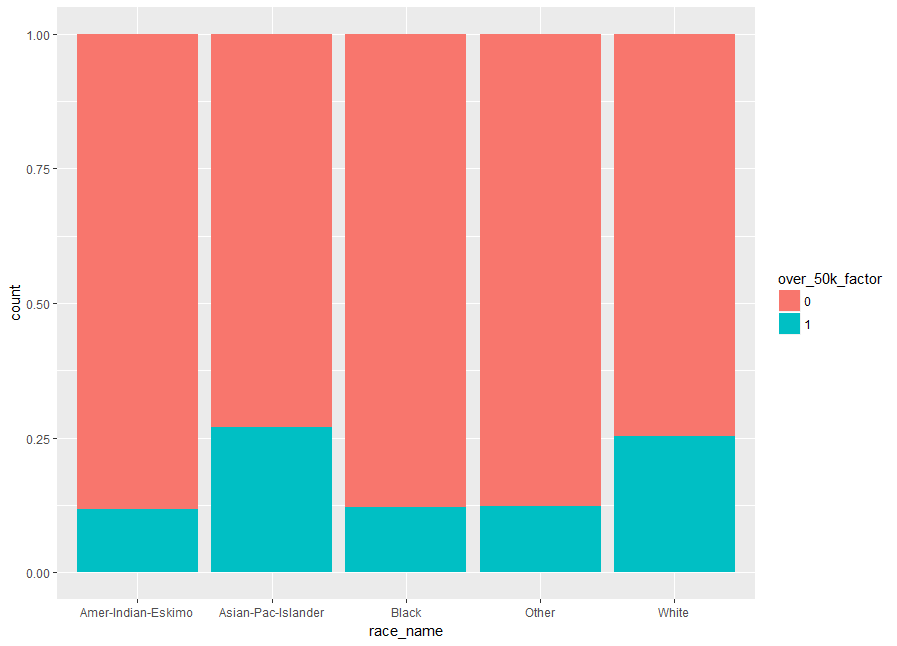


Figure : Percentage of Those Who Made Over $50,000/year by Race

Another interesting finding was visualizing the relationship between how many years of education a person has and the percentage of those in that category that make over $50,000/year. The percentage increases as education increases, which is to be expected. However, it is not linear. The green line represents a cubic relationship, which seems to fit the data more accurately.

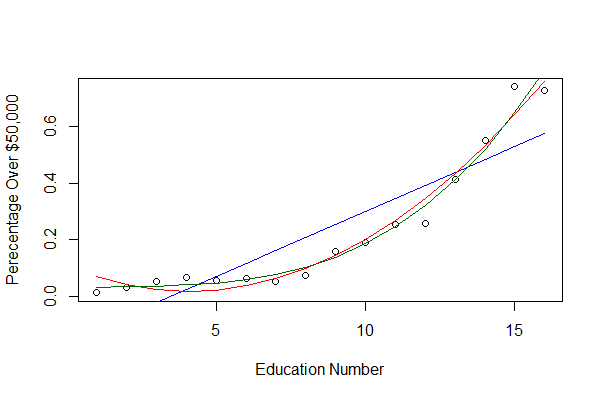
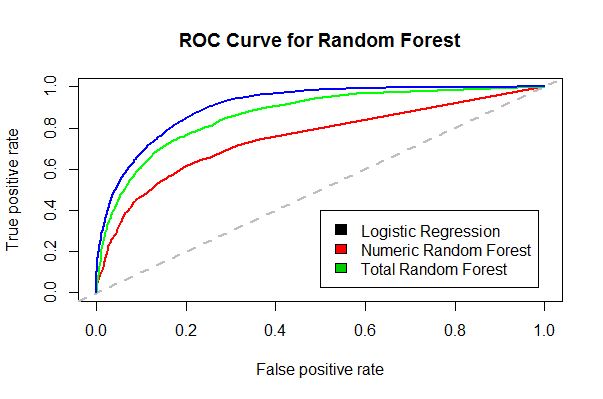


Figure : Relationship Between Years of Education and Percentage of Those Who Made Over $50,000/year

**Model Building**

As stated earlier, two types of models were explored in order to best predict if a person makes over $50,000/year. A ROC curve is used in order to compare the accuracy of models across multiple models. The closer the line is to the upper left-hand corner, the more accurate. From this it is clear that the Logistic Regression provided the most accurate model.



**Conclusions and Recommendations**

The final model created with logistic regression took all variables in and provided an accuracy of 84.8%. If a simpler model is desired, there are a few variables, such as sex, country, and race, that could be dropped while not sacrificing a large amount of accuracy. However, there are a few variables that should not be dropped because of their effect on the model. For example, the marital status has a large impact on the model. Those who are married are more likely to make over $50,000/year. Also, number of hours a week the person works should also not be dropped, which makes sense. The more hours you work, the more likely you are to make over $50,000/year.