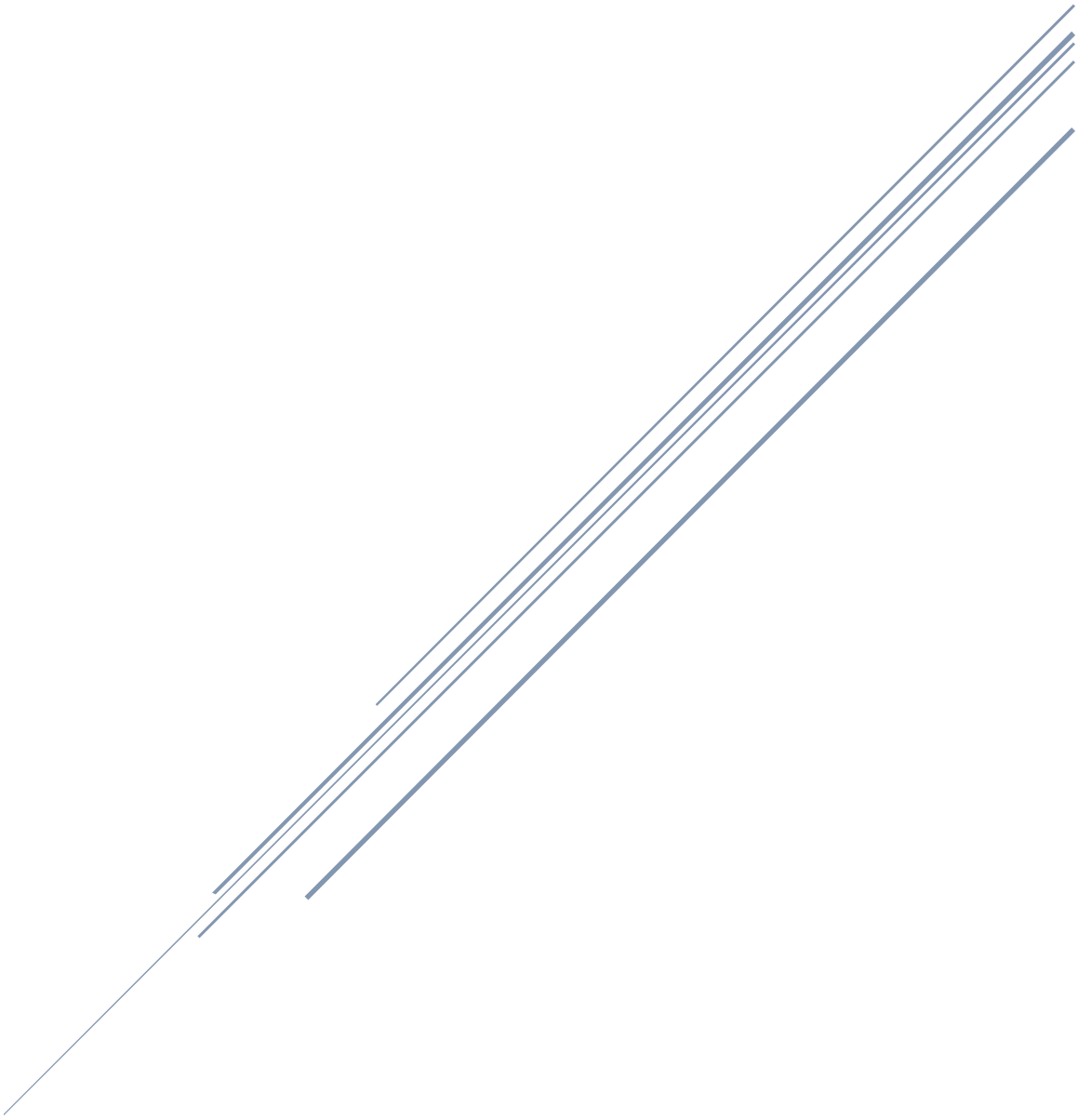
PREDICTING SALARY LEVEL FROM US CENSUS DATA

RTI APPLICATION – DATA SCIENTIST (JOB ID: 16715)

Sarah Mohamed



January 26, 2016

Executive Summary

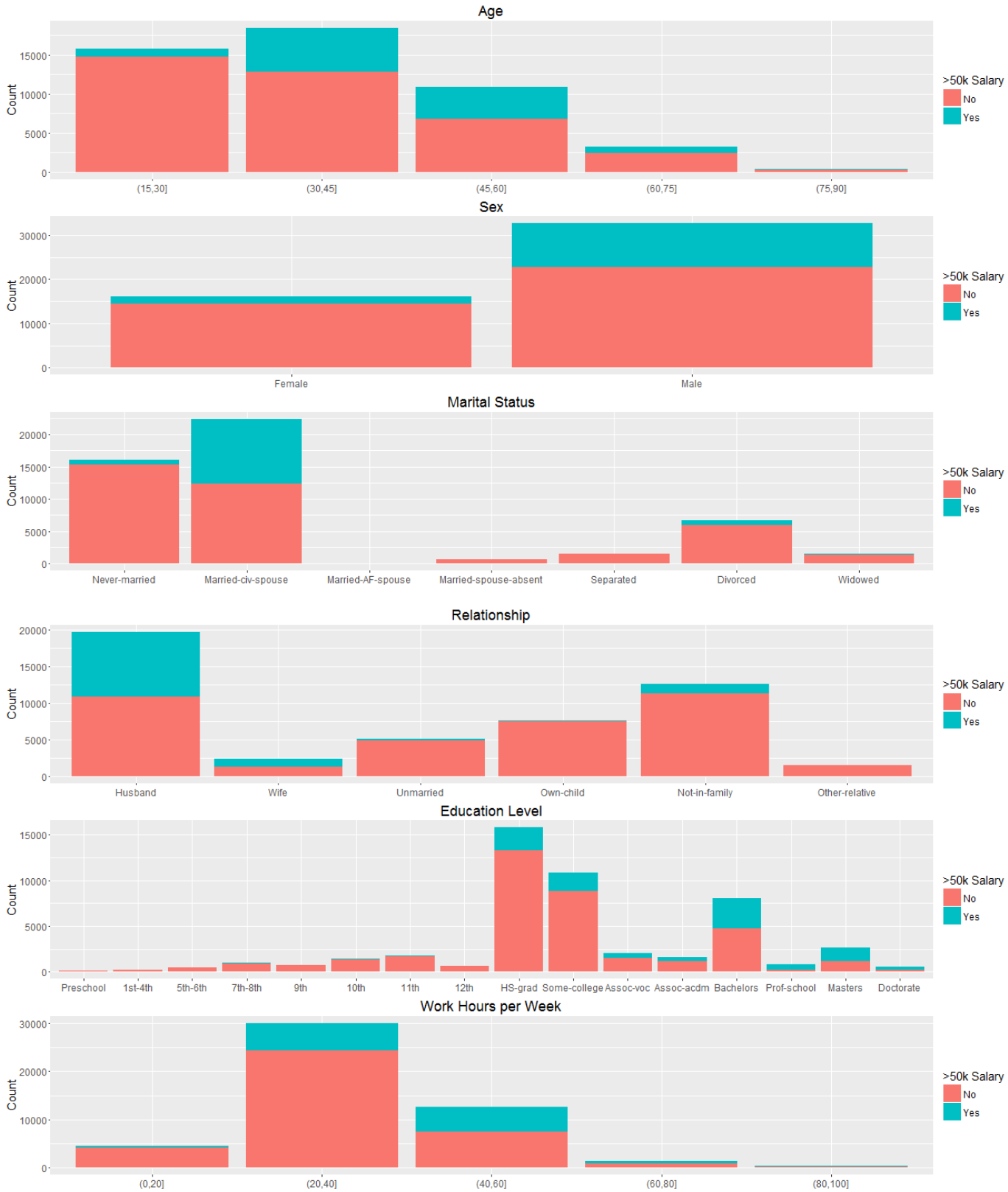
Population attributes from the US Census are used to build a model that predicts if an individual has a salary greater than $50,000. A sample of 48842 personnel are used for this analysis with 24% of the sample earning over $50,000. Using R software, a logistic regression model is built with attributes that are found statistically significant at an alpha of 0.001. These attributes are age, sex, marital status, relationship, education, and work hours per week. To visualize the relationship of these attributes, Figure 1 shows the salary level of personnel in each category of these population characteristics.

The model has a c-statistic of 0.88 and a misclassification rate of 24% when tested against both validation and testing hold out samples. A cutoff value of 0.21 is used for this model, indicating that when the model results in a probability of at least 21%, that personnel is categorized as making greater than $50,000. This cutoff value is determined by maximizing both the sensitivity and specificity of the model performance. Using this methodology minimizes the overall false positives and false negatives produced by the model.

In preparation for model development, the data is explored for any insufficiencies that would affect variable performance in the model. As a result of this process, the variables workclass and country are altered to correct for quasi-complete separation. Also, all continuous variables are binned so that assumptions of logit linearity are not a concern. Finally, after the variables are graphed, capital loss and capital gain are identified as highly skewed variables. However, neither variable enter the final model since capital loss is found to be insignificant and capital gain is dismissed due to quasi-complete separation.

Further analysis can be performed on this model to identify significant interactions between the variables. Also, other modeling techniques including neural networks and decision trees can also be developed and compared to the logistic regression in this analysis.

Figure 1: Greater than $50,000 Salary vs. Population Attributes



SQL Query Code

```
CREATE TABLE records_flat AS

SELECT r.id, r.age, w.name as Workclass, e.name as Education_level, r.education_num, m.name as Marital_Status, o.name as Occupation, re.name as Relationship, ra.name as Race, s.name as Sex, r.capital_gain, r.capital_loss, r.hours_week, c.name as Country, r.over_50k

FROM records as r, countries as c, education_levels as e, marital_statuses as m, occupations as o, races as ra, relationships as re, sexes as s, workclasses as w

WHERE r.country_id = c.id and r.education_level_id=e.id and r.marital_status_id=m.id and r.occupation_id=o.id and r.race_id = ra.id and r.relationship_id=re.id and r.sex_id=s.id and r.workclass_id=w.id
```