1. **Problem Statement**

Improve accuracy of self reported income for prospective clients of a credit card company. Need to identify key demographic variables and corresponding predictive model that segments the prospect into high income / low income with a threshold between Low/High at $50,000 ( Threshold driven by best available labeled reliable data source)

*Proposal*

Predict individual salary based on demographic and employment data collected as part of the 1994/95 survey and leverage the predicted salary to compare against reported data.

1. **Client**

Client is a credit card company that wants to combine self reported income details by their clients against those predicted based on client demographics collected from various sources. They would use this analysis to understand the most critical data points needed for a reasonable accurate income prediction and then be able to combine self reported and predicted incomes for decisions like offer targeting and credit line assignment. The goal is to give higher lines to high earning individuals and minimize self reporting errors or system gaming

1. **Data Sourcing**

I will be using a 1994/95 census data set with demographic and employment variables.

This data set contains weighted census data extracted from the 1994 and 1995 [Current Population Surveys](http://www.census.gov/cps/methodology/techdocs.html) conducted by the U.S. Census Bureau. The data contains 41 demographic and employment related variables.

The instance weight indicates the number of people in the population that each record represents due to stratified sampling. To do real analysis and derive conclusions, this field must be used. This attribute should \*not\* be used in the classifiers.

<http://archive.ics.uci.edu/ml/machine-learning-databases/census-income-mld/census-income.data.html>

1. **Data Definition**

|  |  |  |
| --- | --- | --- |
| **Categorical Variables** | | |
| '\_class\_of\_worker' | '\_reason\_for\_unemployment' | '\_migration\_prev\_res\_in\_sunbelt' |
| '\_industry\_code' | '\_full\_or\_part\_time\_employment\_stat' | '\_family\_members\_under\_18' |
| '\_occupation\_code' | '\_tax\_filer\_status' | '\_country\_of\_birth\_father' |
| '\_education' | '\_region\_of\_previous\_residence' | '\_country\_of\_birth\_mother' |
| '\_enrolled\_in\_edu\_inst\_last\_wk' | '\_state\_of\_previous\_residence' | '\_country\_of\_birth\_self' |
| '\_marital\_status' | '\_detailed\_household\_and\_family\_stat' | '\_citizenship' |
| '\_major\_industry\_code' | '\_detailed\_household\_summary\_in\_household' | '\_own\_business\_or\_self\_employed' |
| '\_major\_occupation\_code' | '\_migration\_code-change\_in\_msa' | 'fill\_inc\_questionnaire' |
| '\_mace' | '\_migration\_code-change\_in\_reg' | '\_veterans\_benefits' |
| '\_hispanic\_Origin' | '\_migration\_code-move\_within\_reg' | '\_weeks\_worked\_in\_year' |
| '\_sex' | '\_live\_in\_this\_house\_1\_year\_ago' | 'year' |
| '\_member\_of\_a\_labor\_union' |  |  |

|  |
| --- |
| **Continuous Variables** |
| 'age' |
| '\_wage\_per\_hour' |
| '\_capital\_losses' |
| '\_capital\_gains' |
| '\_divdends\_from\_stocks' |
| '\_num\_persons\_worked\_for\_employer' |
|  |
|  |
|  |
|  |
|  |
|  |

1. **Key Data Wrangling Steps for Income Census dataset**

The existing data set has 42 columns

1 target column (\_Income\_category)

1 column specifying instance weights – Indicates the population colume the row represents. Not to be used as a model input but as weights

31 Categorical columns

8 Continuous columns

|  |  |  |  |
| --- | --- | --- | --- |
| **Observation** | **Description** | **Columns Impacted** | **Discussion and Actions taken** |
| **Handling Missing Values** | Following 8 columns have missing values , represented by ‘?’ | \_region\_of\_previous\_residence  \_migration\_code-change\_in\_msa  \_migration\_code-change\_in\_reg  \_migration\_code-move\_within\_reg  \_migration\_prev\_res\_in\_sunbelt  \_country\_of\_birth\_father  \_country\_of\_birth\_mother  \_country\_of\_birth\_self | I’d keep the missing value as **‘?’** for now. There is no obvious reason for missing values beyond the fact that responders forgot to enter those columns, ex. Info. Around previous residence and birth country seems missing.  **Action:** Keep the value ‘?’ as of now to review any not so obvious relationships between missing values and other parameters |
| **‘Not in Universe’ values observed** | These 14 columns have values that cannot be categorized and are bucketed into a ‘Not in Universe’ category. | \_major\_industry\_code',  '\_major\_occupation\_code',  '\_reason\_for\_unemployment’, '\_region\_of\_previous\_residence',  '\_state\_of\_previous\_residence',  '\_migration\_code-change\_in\_msa',  '\_migration\_code-change\_in\_reg',  '\_migration\_codemove\_within\_reg',  '\_migration\_prev\_res\_in\_sunbelt',  '\_family\_members\_under\_18',  '\_country\_of\_birth\_father',  '\_country\_of\_birth\_mother',  '\_country\_of\_birth\_self',  'fill\_inc\_questionnaire' | The source doesn’t provide any reasoning for this. Most likely these are driven by incomplete answers or responses that are skipped or not valid. NO obvious correlation observed between ‘Not in Universe’ values across columns, based on simple cross tab analysis  **Action:**  Plan to keep these as it is to analyze any not so obvious relations and allow for appropriate modifications |
| **Outliers** |  | No outliers observed – Data is weighted by population count for each row ~1800 counts associated with each row. No obvious outlier row observed | NO Action taken |

1. **Initial findings around Income distribution**
2. High Income earners are mostly > 20 years of age and concentrated between 37 and 62 years of age, which are expected to be the core earning period

Low income earners are more spread out due to low earning teenage population, as shown in Fig 1 below

|  |  |
| --- | --- |
| **Fig1 : Income vs Age** | **Fig2 : Income vs Employer Size** |
|  |  |

1. Fig 2 indicates that most high earners are concentrated in larger firms.

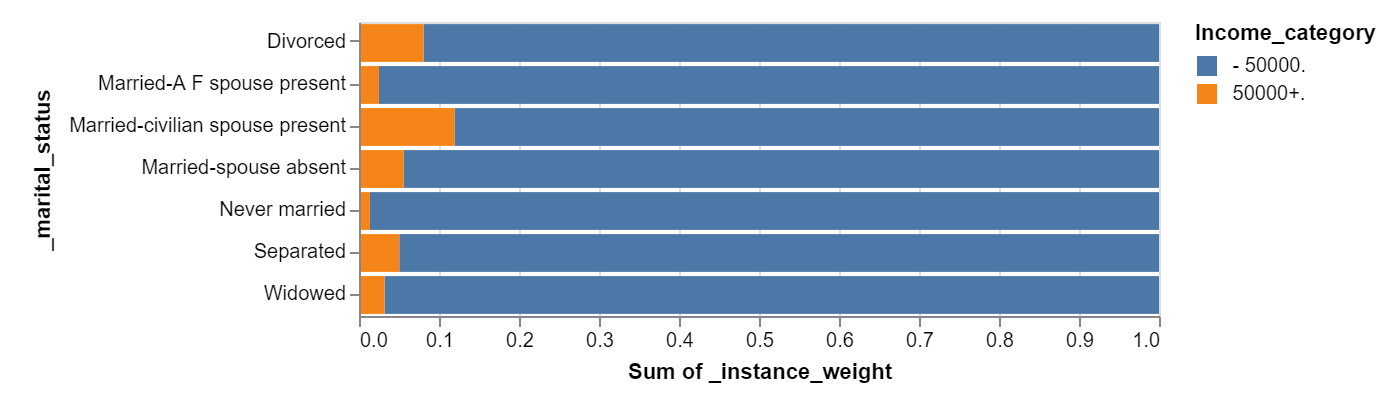
**Fig 3a): Heat map of % High earners by birth parents’ country of origin**

**Fig 3b) Income distribution based on either/both/No parent with US country of origin**

|  |  |
| --- | --- |
| **3 a)** | **3 b)** |
|  |  |

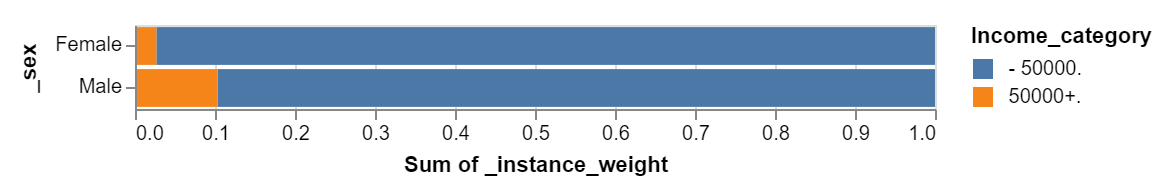
1. Fig 3 a) Some of the dark red cells tend to be driven by lower population in the cell (low denominator)
2. The data agrees with common beliefs around mating patterns – ex. First generation Asian-Indians dont seem to have non Indian partners, which makes sense in 1994-95 (Period of data), given conservative views. In a similar vein English/French/German nationals seem to have partners amongst each other
3. Fig 3 a) above we see some dark areas of higher concentration of high earners within population where both parents have a non-US country of origin.
4. Fig 3 b) indicates that income potential is relatively high when both / either parent is from US. Need to analyze this further as we try to model the relationship with target

**Fig 4: Income distribution by Marital Status**



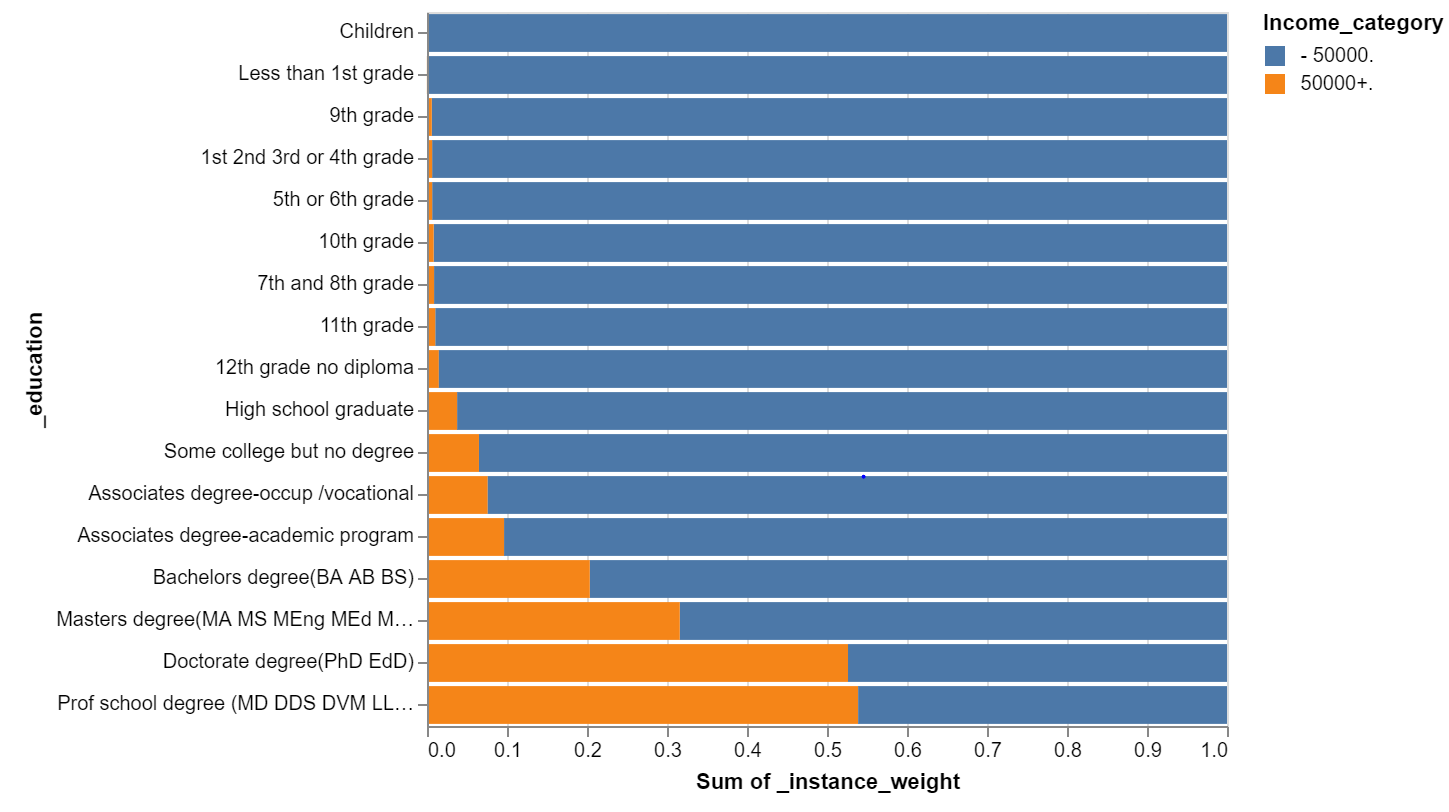
1. Marriage is good for earning potential: Married couples, where both spouses are present tend to earn more compared to divorced or married with a single spouse. Division of household responsibilities possibly allows more time for career resulting in higher income potential

**Fig 5: Income by Gender**



1. There is irrefutable evidence in Fig 5 that men tend to earn more than women

**Fig 6: Income by Education Level**



1. As expected, level of education has a significant impact on earnings potential, esp. a significant jump if you get a Bachelor’s vs an Associated degree, a Masters vs Bachelors and Doctorate vs Masters