Business Analytics – Project

Project Title – To predict if income of an individual exceeds $50k given the data of a set of individuals having details such as age, gender, qualification, working class, marital status, occupation, race (white, black, asian etc.) hours spent per week and an attribute indicating whether income exceeds $50k or not.

Project Description – This project is based on the data sets provided in UCI and Kaggle repositories for business analytics projects. This is classification problem using supervised learning method of modelling. It can be used for predictive analytics purpose to determine if income of a certain individual exceeds a threshold.
This is related to a problem sometimes faced by government agencies in determining income of individuals to calculate tax liability and to confirm whether any individual is out of tax net or to which tax slab an individual belongs. We could identify few regression problems as well for which this same data set could be used. E.g. We could determine if there was any relationship between race and income to figure out if there was any racial bias for income. But, as far as this project is concerned we would limit ourselves to build a classification model to determine if income is above or below $50k

Objective – To prepare a predictive analytics model to determine if income is below or above $50k threshold for a new set of similar data presented to the model

Scope

- Problem type – classification
- Original data set contains 48842 observations. Out of that, for 2800 observations, data is missing for 'working class' and 'occupation' attributes. All such observations would be excluded for modelling given that the remaining dataset would still have approximately 46000 observations
- For 857 observations data is missing for attribute 'native-country'. These observations would be also be excluded from modelling
- As per the dataset details as UCI data is missing for 2 workclass, 7 occupation and 14 native countries. All such observations with missing values would be out of scope of modelling
- Attribute 'fnlwgt' is excluded from the scope as we could not find the meaning and significance of this attribute in the context of this problem

Method - Logistic regression method for prediction

Data Snapshot
age – Age of an individual. unit – years, type – continuous

workclass - working class of an individual. Type – categorical, set: *(Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked)*

education - Qualification of an individual. Type – categorical, set: *(Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool).*

education-num - number to each category of educational qualification. Type –continuous

marital-status - Type – categorical, set:*(Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse)*

occupation - Type – categorical, set: (*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, Armed-Forces.)*

relationship - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race – Racial identity. Type – categorical, set: (*White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black)*

sex: Type – categorical, *set: (Female, Male)*

capital-gain - Unit - USD, Type – continuous

capital-loss - Unit - USD, Type – continuous

hours-per-week - Effort spent at work during a working week. Unit – hours, Type – continuous

native-country – Country of origin, Type – categorical, set: (*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)*

Below is the snapshot of actual data. ‘fnlwgt’ attribute has been excluded

age	**workclass**	**education**	**educational-num**	**marital-status**	**occupation**	**relationship**	**race**	**gender**	**capital-gain**	**capital-loss**	**hours-per-week**	**native-country**	**income**
17	Private	10th	6	Never-married	Machine-op-inspct	Not-in-family	White	Male	0	0	40	United-States	<=50K
39	Private	Some-college	10	Divorced	Sales	Not-in-family	White	Male	0	0	40	United-States	<=50K
26	Private	Masters	14	Never-married	Exec-managerial	Not-in-family	White	Female	0	0	40	United-States	<=50K
21	Private	Some-college	10	Separated	Handlers-cleaners	Not-in-family	White	Female	0	0	40	United-States	<=50K
20	Private	Some-college	10	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	40	United-States	<=50K
23	Private	Bachelors	13	Never-married	Exec-managerial	Not-in-family	White	Male	0	0	40	United-States	<=50K
24	Private	Some-college	10	Separated	Other-service	Not-in-family	White	Male	0	1876	40	United-States	<=50K
24	Federal-gov	Some-college	10	Never-married	Armed-Forces	Not-in-family	White	Male	0	0	40	United-States	<=50K
59	Federal-gov	Bachelors	13	Divorced	Adm-clerical	Not-in-family	White	Male	0	0	40	United-States	<=50K
34	Private	Masters	14	Never-married	Other-service	Not-in-family	Amer-Indian-Eskimo	Male	0	0	40	United-States	<=50K
36	Private	10th	6	Separated	Machine-op-inspct	Not-in-family	White	Female	0	0	40	United-States	<=50K

41	Private	Some-college	10	Divorced	Tech-support	Not-in-family	White	Female	0	0	40	United-States	<=50K
38	Self-emp-inc	Bachelors	13	Divorced	Exec-managerial	Not-in-family	White	Male	0	0	40	United-States	<=50K
34	Private	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Male	0	0	40	United-States	<=50K
26	Private	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	40	United-States	<=50K
26	Private	Some-college	10	Never-married	Sales	Not-in-family	White	Female	0	0	40	United-States	<=50K
62	Federal-gov	Some-college	10	Divorced	Tech-support	Not-in-family	Black	Male	4650	0	40	United-States	<=50K
33	Private	Bachelors	13	Never-married	Prof-specialty	Not-in-family	White	Female	0	0	40	Canada	<=50K
19	Private	Some-college	10	Never-married	Adm-clerical	Not-in-family	White	Male	2176	0	35	Germany	<=50K
30	Private	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Male	0	0	40	India	<=50K
36	Private	Bachelors	13	Divorced	Sales	Not-in-family	White	Female	0	0	40	India	<=50K
49	Private	Bachelors	13	Never-married	Tech-support	Not-in-family	White	Female	0	1564	40	Canada	>50K
22	Private	Bachelors	13	Never-married	Other-service	Not-in-family	White	Female	0	0	35	Canada	<=50K
26	Private	Some-college	10	Never-married	Protective-serv	Not-in-family	Black	Male	0	0	55	Philippines	<=50K
43	Private	Masters	14	Divorced	Prof-specialty	Not-in-family	White	Male	0	0	40	Canada	<=50K

Data sources (adult.csv)

- UCI dataset web page **-** http://archive.ics.uci.edu/ml/datasets/Adult
- Data set csv file download page - https://www.kaggle.com/wenruliu/adult-income-dataset