Project Report

Group Members – Sakshi Jain
Chinmay Bhat
Shri Lalana Reddy
Trapti Kandelwal
Ram Warutkar

Topic - Texas Census Data Analysis w.r.t Prediction of Individual Annual Income.

Project Overview – Income is often a useful indicator of a society's quality of life and wealth. Household income is a common cited measure of individual wealth. For e.g. Economists use household income to draw a host of conclusions about the economic health of a given area or population. Per capita income measures the average income earned by each person in a given area. Monitoring HH income is important for government, donors, researchers, and others because increasing rural household income is a primary objective for achieving many development goals, including reducing poverty, hunger, and food and nutrition insecurity. So, we here as a team, tried to develop a machine learning model to predict the income.

Business Understanding Researchers, policy makers and administrators routinely face the problem of selecting an observable indicator of welfare from cross-sectional data. These indicators are expected to convey information about the welfare of households well beyond the survey period. The leading practical indicator in this respect is income.

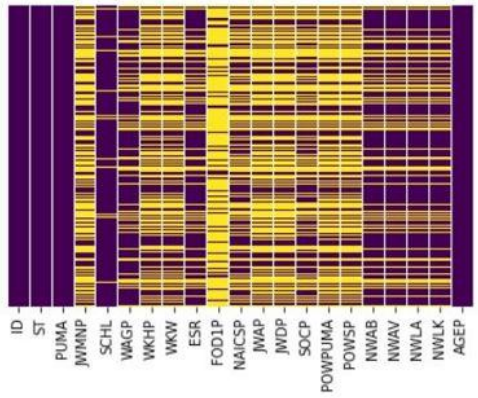
Problem Statement - To find individual annual income to help policy makers of the country with the public investment programs. Develop robust regression model to predict the income. Explanatory analysis to find indicators effecting income.

Data Understanding - The data file contains data from the 2015 American Community Survey. It is limited to Texas, and contains a subset of the total survey variables. The full columns are –

1. ID-Identification
2. ST-State code
3. PUMA-Public use microdata area code
4. JWMNP-Travel time to work
5. SCHL-Educational attainment
6. WAGP-Wages
7. WKHP-Usual hours worked per week past 12 months
8. WKW -Weeks worked during past 12 months
9. ESR-Employment status recode
10. FOD1P-Recoded field of degree
11. NAICSP-NAICS industry recode for 2013
12. JWAP-Time of arrival at work - hr&min
13. JWDP-Time of departure at work - hr&min
14. SOCP-SOC occupation code
15. POWPUMA-Place of work
16. POWSP-Place of work –state of foreign country recode
17. NWAB-Temporary absence from work
18. NWAV-Available for work
19. NWLA-On layoff from work
20. NWLK-Looking for work
21. AGEP-Age

There are 21 columns of which 15 float, 4 int and 2 object and 259224 rows and total observations are 5443704. There are 1752554 null values that is 32.19% of whole data.

Null Value Plot



Fields in the dataset can be group according to the type of information they convey.

Area / Location – ST(int64), PUMA(int64), POWPUMA(float64), POWSP(float64)

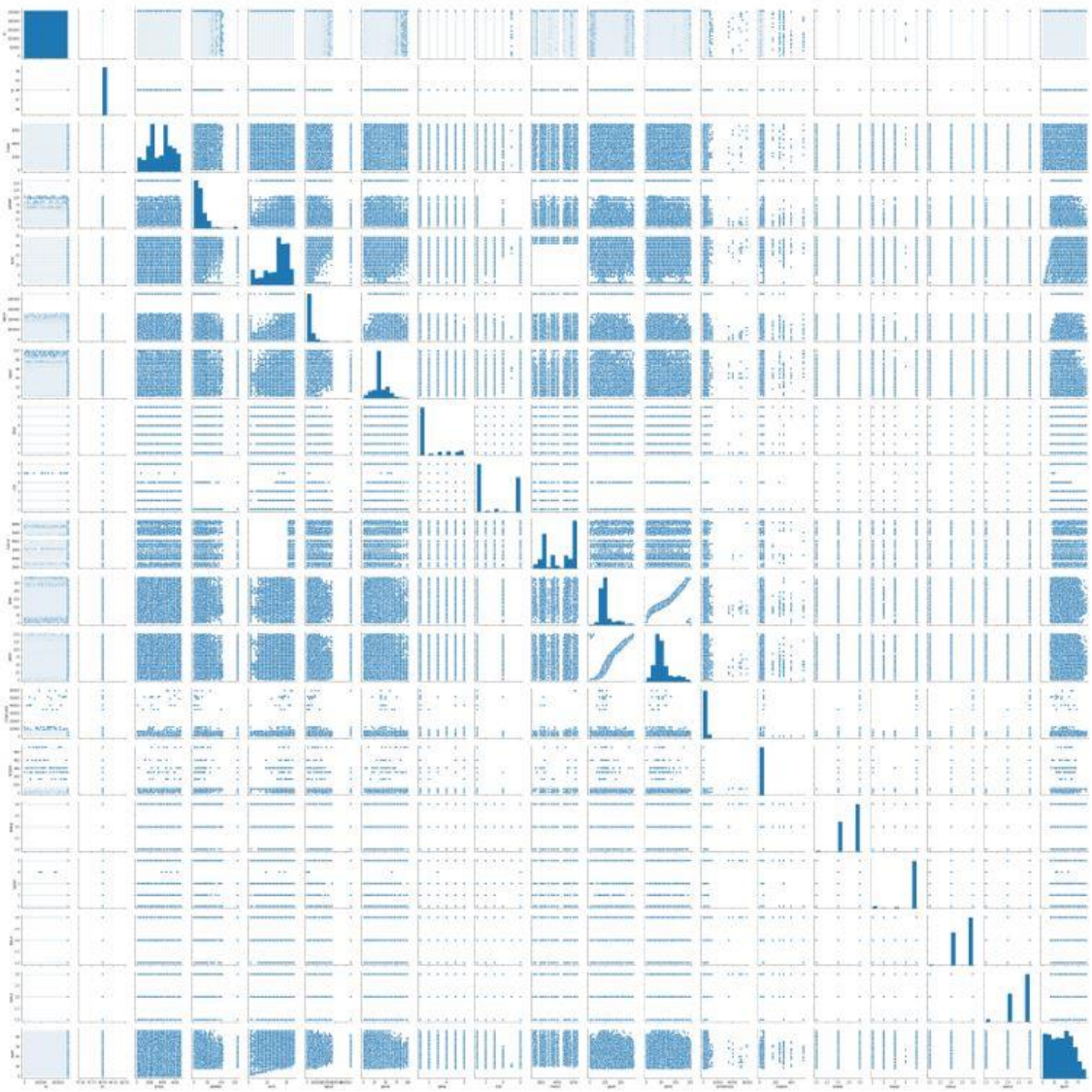
Time / Duration – JWMNP(float64), WKHP(float64), WKW(float64), JWAP(float64), JWDP(float64)

Employability Status – ESR(float64), NWAB(float64), NWAV(float64), NWLA(float64), NWLK(float64)

Qualification – FOD1P(float64), SCHIL(float64)

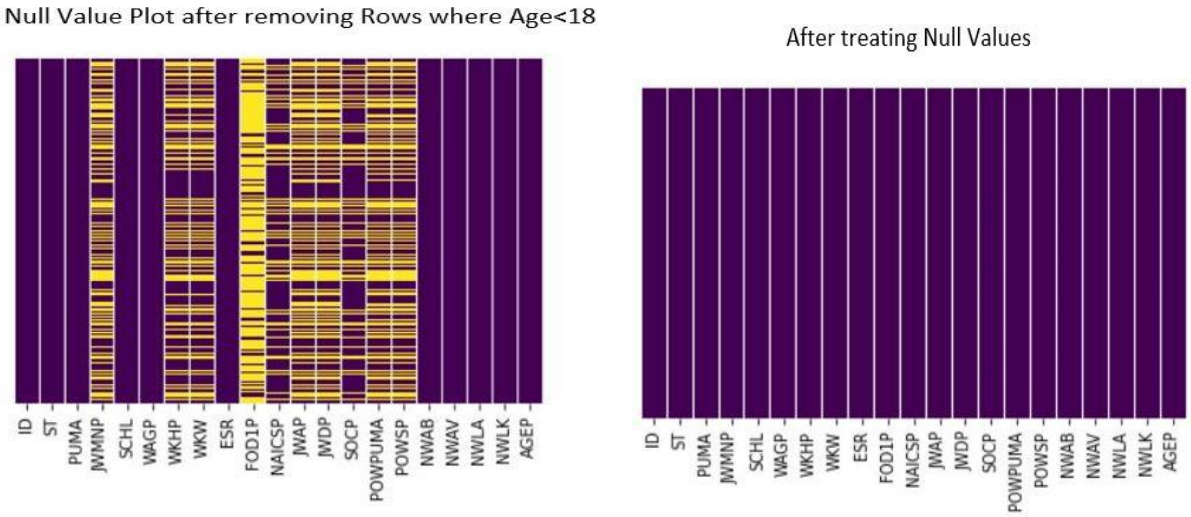
Others – AGEP(int64), SOCP(object), NAICSP(object),ID(int64),WAGP(float64)

Pair Plot



Data Preparation (File Name - 1_Data_Preparation.ipynb)

1. A quick observation on Age column in dataset indicates that there are entries which are less that 18 years, and the minimum employability age in Texas is 18 years (Source -Internet). So, the rows with age less than 18 years is potentially false data and should be removed. There are 64741 columns were age is less than 18 is 24.97 % of the total observations. The shape of the dataset is reduced to (194483,21) and the null values are reduced to 803614 i.e. 54.14 % decrease.
According to the documentation of the dataset JWNMP has N/A value where not a worker or worker who worked at home. Similar is the case with SCHL, WKHP, WKW, ESR, FOD1P, NAICSP, JWDP, SOCP, POWPUMA, POWSP

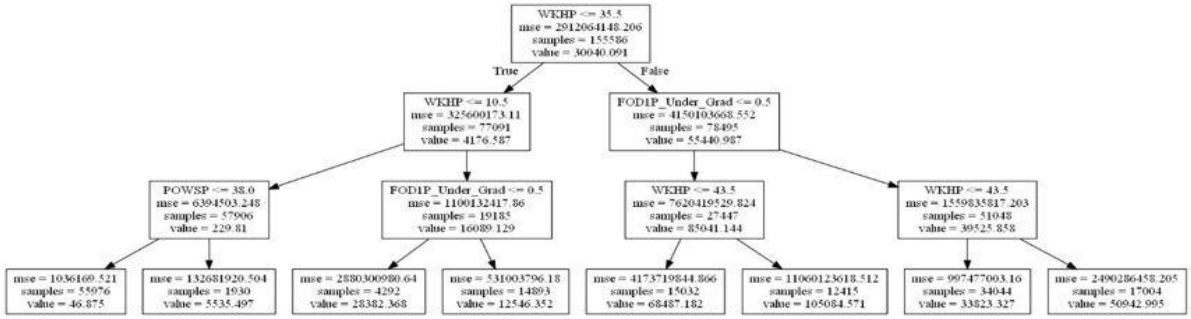


2. After treating Null Values now in the column of SCHL (Educational attainment) it is observed that the education level can be grouped together as Pre School, School, High School, Graduation, College, Post-Graduation, No Schooling. Grouping them according to their designated codes. Similarly, the column FOD1P - Recoded field of degree has a large number of levels, which is also grouped to reduce to, two categories Under Graduate and Graduate and above. In the same fashion NAICSP column is also treated. Where it can be grouped by the type of industry (e.g. Agriculture, Manufacturing etc.). In the similar fashion JWAP, JWDP, POWPUMA columns are grouped.
In JWAP column each time interval is separated by 15 minutes is so, four quarters of 6 hours are made (Morning 12 AM to 12 PM, Evening 12 PM to 11:59 PM), similar with JWDP. In Column POWPUMA three categories are made (Assigned POW, Not Worker or Under16, Not Work in USA or Puerto Rico) and grouped according the assigned codes.

Model Building –

1. Decision Tree – (File Name - 2_Regression_Decision_Tree.ipynb)
1. The first task in building regression tree was to identify Categorical variables and converting it to Numerical variables. There 7 categorical fields in the dataset SCHL, FOD1P, NAICP, JWAP, JWDP SOCP, POWPUMA, one hot-encoding is used to do the transition. Simple decision tree with max depth 3 yields test accuracy of 33.93% and train accuracy of 34.46 which indicates there is less chance of overfitting.

Simple Decision Tree Max Depth 3



Further to improve the model performance various Decision Tree ensemble techniques can be used.

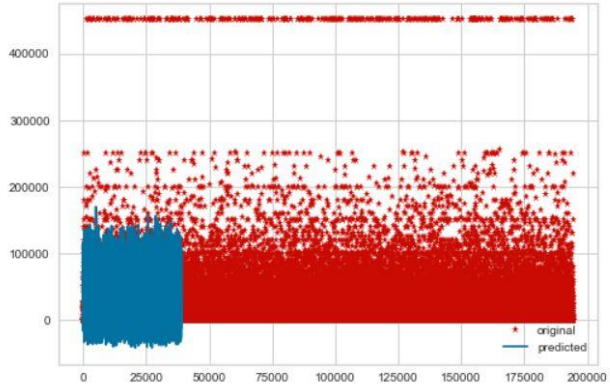
1.1 Bagging

1.2 Random Forest

1.3 Boosting Techniques

2. Support Vector Machine

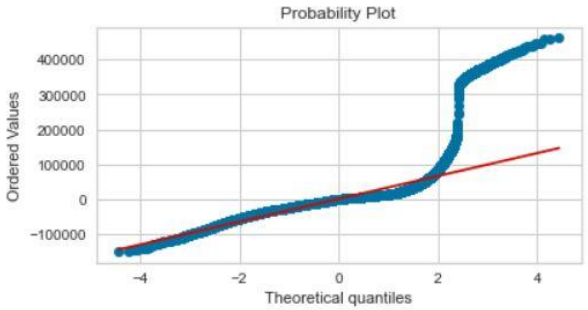
3. Linear Regression (File Name - 4_Regression_Linear_Regession.ipynb)
The first task in building regression tree was to identify Categorical variables and converting it to Numerical variables. There 7 categorical fields in the dataset SCHL, FOD1P, NAICP, JWAP, JWDP SOCP, POWPUMA, one hot-encoding is used to do the transition. In linear regression the test r square values is 36.711% and train r_square is 37.12 which indicated that there is less chance of over fitting but the explain ability of the model is less than 50%



Checking for assumptions –

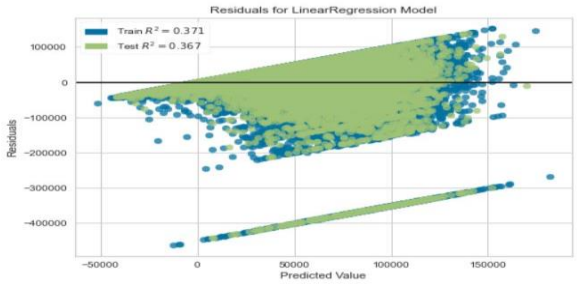
1. Normality

Since the line is not following the straight line it indicated that the distribution may not be normal



2. Residual Plot

Residual plot also indicates that the variation is not uniform



DECISION TREE

R_Square	Train RMSE	Test RMSE
50 per cent	42042	37841

RANDOM FOREST

R_Square	Train RMSE	Test RMSE
49 per cent	39763	38297

BAGGING

R_Square	Train RMSE	Test RMSE
91 per cent	15928	41607

ADAPTIVE BOOSTING

R_Square	Train RMSE	Test RMSE
31 per cent	44748	44500

Gradient Boosting

R_Square	Train RMSE	Test RMSE
37 per cent	42578	42565