**Name – Arpit Mohanani**

**Student ID – 300337709**

**Predicting Income Range of an Adult**

* 1. **Introduction and discovery**

In this project we will be focusing on predicting whether an individual's income is more than $50k per year or not using various features like age, education qualifications, marital status, occupation and more.

Using all these features we will predict the category to which the individual belongs i.e. either more than 50k per year or less than 50k per year. We will use various regression models in this project to achieve our goal.

* 1. **Data Preparation**

The dataset includes data of various adults like their age, education qualifications, marital status, occupation and more.

Link to the dataset - <https://archive-beta.ics.uci.edu/ml/datasets/adult>

In data processing, there were around 4000 rows with atleast 1 null value. So we had to remove them. The columns including ‘race’ and ‘relationship’ were also dropped because they were irrelevant to our target. We also had to reduce unique values in various columns in the dataset in order to generalize few features before moving to data modeling.

* 1. **Model Planning**

In the dataframe, we had to create Dummy values first for the categorical columns.

4 different feature models were created using following selection techniques:

**Manual Selection**: In this selection, we selected those 5 features which are having highest correlation with the target.

**Variance Threshold Selection**: Using this selection method, we selected those features which were having variance of more than 0.6.

**Select K Best**: Using this selection method, we selected top 10 best features out of all.

**Random forest feature Selection**: Features selected using this selection method is based on how well the feature improves the purity of the node. Using this method we can get features which are more important than others.

* 1. **Model Implementation**

As mentioned above, I created 4 feature selection models. Before performing regression, transformation and scaling was required to be performed on the dataframe.

I created a pipeline which first transforms the dataframe using Polynomial Features of degree 2. Then the result will fit in a scalar function. I used Standard Scalar for scaling the data. After that finally, regression is performed using 9 different regressors.

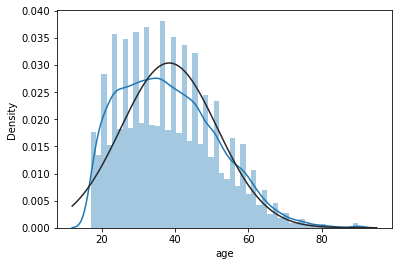
So, there were 4 different dataframes after feature selection and every dataframe was performed with 9 different regressors. So, in the end we got 36 different models with their R2 and RMSE values.

In my case, the best model that I got was with SelectKBest as the feature selection method with CatBoost Regressor with least RMSE value and high R2 value.

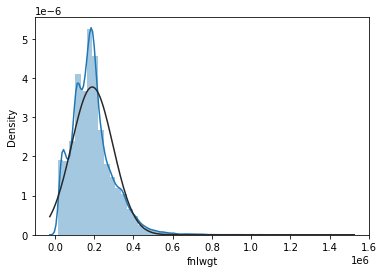
* 1. **Results Interpretation and Implications**

**Univariate Analysis**

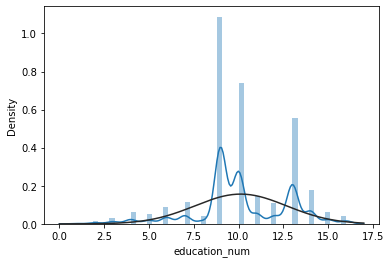
* Age



* Fnlwgt

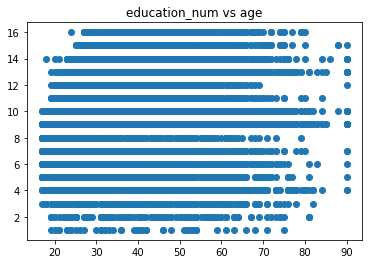


* Education\_num

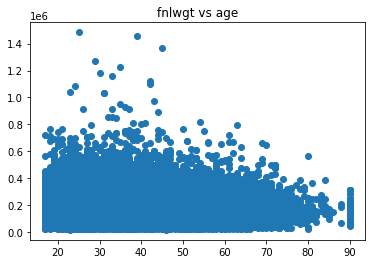


**Multivariate Analysis**

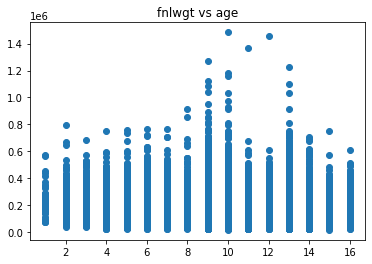
* Education\_num vs Age



* Fnlwgt vs Age

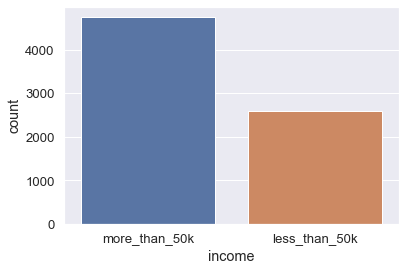


* Fnlwgt vs Education\_num

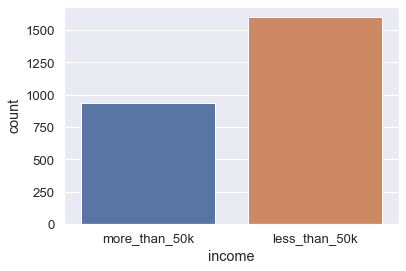


**Count Plots**

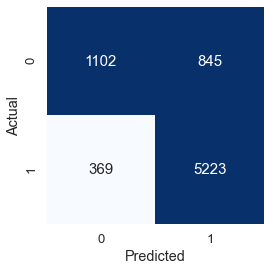
* Age > 75 percentile



* Education\_num > 75 percentile



Best Regression Model to Predict Income:



* 1. **Concluding Remarks**

- Unmarried individuals might have more time for self-development and improve their skills in order to earn more money.

- In most cases education is directly proportional to the earnings. The better the education a person gets the more he/she might earn.

- Age is also a major factor to consider. With age comes experience and more experienced person achieves more in life.

- The more the person works the more he/she earns. So number of hours worked per week is directly proportional to the income of the person and is a strong factor to be considered.