### **Data Source**

This data set contains weighted census data extracted from the 1994 and 1995. Each instance contains 40 demographic and employment-related variables as well as an *instance weight* and a label for each observation indicating population with annual income <50k and > 50k. Total there are 199,523 rows.

1. **Data pre-processing:**

To achieve better results from the model, the format of the data has to be in a proper manner. The Data pre processing includes two concepts such as **Data Cleaning**and **Feature Engineering.**

**2.1 Data Cleaning:**

Category Reduction: There are some columns in the data set which has many categories and

we reduced the categories for a better modelling. The columns are: marital stat..

Missing values: Check for missing, null and na values.

Duplicate Values:Removed duplicate values from the data.

### ****2.2 Outlier Detection:****

### Detecting outliers in the data is a extremely important step in data preprocessing. The presence of outliers can often skew results.

### ****2.3 Dummy variables:****

Encoding all of the categorical independent variables as dummy variables allows easy interpretation , calculation and increases the stability and significance of the model coefficients.

### ****2.4 Feature Scaling:****

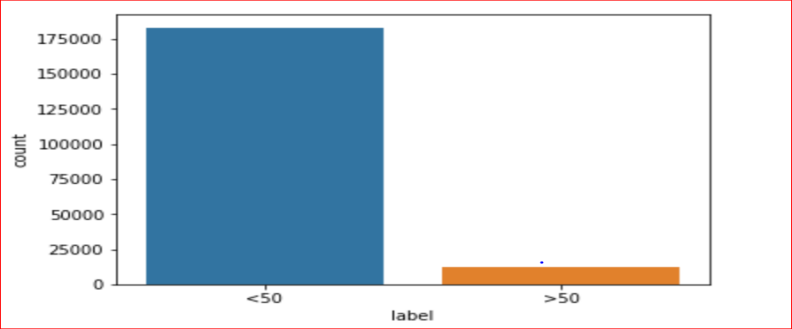
### After applying a natural logarithm scaling to the data, the distribution of each feature should appear much more normal. For any pairs of features correlation should persist after transformation.

**2.5 Standardize Data:**

Standardization is a useful technique to transform attributes with a Gaussian distribution and having different means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1 (μ=0 and σ=1).

**2.6 Class Imbalance:**

As the response variable(label) is a binary categorical variable (<50k or >50k), we need to make sure the training data has the approximately equal proportion of each class. But in our data set there were 94% customers with annual income <50k and 6% customers with annual income >50k.



Here we used up-sampling i:e duplicating observations with replacement from the minority class i:e customers with annual income greater than 50k so that both classes will have nearly same number of observations.

**2.7 Correlation Check:**

To implement logistic regression in our data set, we have to make sure that there should have little or no multicollinearity among the variables.



1. **Model Architecture.**

There are two business problem given.

1. Identifying two groups of people for marketing purposes: People who earn an income of less than $50,000 and those who earn more than $50,000.
2. Basic segmentation model of the people represented in this data set in the context of marketing.

**First Business Requirement:**

**Objective**: Classify customers who earn an income of less than $50,000 and those who earn more than $50,000 from the explanatory variables.

I have used the **Logistic regression** for classifying the income label of customer.

**Logistic Regression**

**What is Logistic Regression:**

Logistic regression is basically a supervised classification algorithm. The target variable(or output), y, can take only discrete values for given set of input features, X. Logistic regression can be classified as binomial, multinomial or ordinal.We will be implementing binomial logistic regression fr our problem.

The core method or heart for Logistic Regression is the **logistic function**. The [logistic function](https://en.wikipedia.org/wiki/Logistic_function), also called the **sigmoid function** was developed by statisticians to describe properties of population growth in ecology. *It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.*

Logistic regression equation is as follow y = e^(b0 + b1\*x) / (1 + e^(b0 + b1\*x)),

where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value x.

**Why Logistic regression?**

This is a classification problem which means we need to classifier users into two groups.

Group1 : with annual income **greater than 50,000**

Group 2: with annual income **less than 50,000.**

Logistic regression can be used to model and solve this kind of binary classification problems. Here we get a probability score that reflects the probability of the occurrence of the event.

Before implementing logistic regression we need to make sure that data is satisfying all the underlying assumptions.

**Logistic Regression Assumptions:**

* Binary logistic regression requires the dependent variable to be binary.
* For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
* Only the meaningful variables should be included.
* The independent variables should be independent of each other. That is, the model should have little or **no multicollinearity**.
* Logistic regression requires quite **large sample sizes**. *Our data set is large enough.*

I have plotted the correlation graph, individual data analysis plot, to ensure that all the assumptions are satisfied.

**Project Implementation (Logistic Regression):**

After some basic data exploration, I found the there is some feature which could predict the income label of the customer more efficiently. The features are:

['age', 'class of worker', 'education', 'wage per hour', 'marital stat', 'enroll in edu inst last wk' , 'major occupation code' , 'race' , 'sex', 'member of a labor union', 'full or part-time employment stat', 'region of previous residence' , 'detailed household summary in household' , 'num persons worked for employer', 'capital gains', 'capital losses', 'family members under 18']

**Step1:** Split the data using train\_test\_split from sklearn. We now split the data into training and test sets. 75% of the data will be used for training and 25% for testing.

**Step2:** Fitting Logistic Regression to the training set data. We have considered the instance weight in the model learning phase.

**Step3:** Then predict new label of unseen data I:e the test data.

**Step4:** Evaluated model using confusion matrix, classification accuracy, ROC curve, classification report etc.

**Business Requirement 2:**

**Objective:** The main objective for this project is to do a **basic segmentation model of the people represented in this data set in the context of marketing.** As a result, we’ll be able to create a digital marketing campaign that will target exactly this market segment. We might be interested to create some different campaigns for the other user segments.

The decision about which variables to use for clustering is a critically important decision that will have a big impact on the clustering solution. Good exploratory research gives us a good sense of what variables may distinguish people or products is critical.

In order to segment of walmart’s customers, we need a way to compare them that is why we are going to use **K-Means clustering** for grouping.

**What is K Mean Clustering:**

*The k-means clustering algorithm is known to be efficient in clustering* ***large data sets****. This algorithm is one of the simplest and the best known unsupervised learning algorithms. It solves the well-known clustering problem. The K-Means algorithm aims to partition a set of objects, based on their attributes/features, into k clusters, where k is a predefined constant.*

**How it works :**

The algorithm defines k centroids, one for each cluster. The centroid of a cluster is formed in such a way that it is closely related, in terms of similarity *( Euclidean distance)* to all objects in that cluster *.* Technically, k-means **minimizes the overall variance**, by assigning each object to the cluster such that the variance is minimized.

**Algorithm:**

Here is the psuedocode for K mean Algorithm:

Initialize the k cluster centers z1, ..., zk

create counters n1, ..., nk -- initialize them to zero

loop

get new data point x

determine the closest center zi to x

update the number of points in that cluster: ni ← ni + 1

update the cluster center: zi ← z i + (1/ni)(x − zi)

end loop

1. Means is just the Expectation Maximization algorithm applied to a particular naive bayes model.

**E-Step**: Coming up with values to hidden variables, based on parameters.

**M-Step**: Coming up with parameters, based on full assignments.

**Why K-Mean:**

There are many statistical methods for clustering and segmentation. Two widely used methods: the Kmeans Clustering Method, and the Hierarchical Clustering Method.

Reason for k-mean:

1. More data sample
2. Continuous explanatory variable
3. No prior label in the data (unsupervised clustering)

\*  Hierarchical Clustering Method works well when the sample size is less.

**Project Implementation:**

## **Select Segmentation Variables**

## After doing some exploratory research and data analysis I found that below listed variables may distinguish people.The attributes included are:

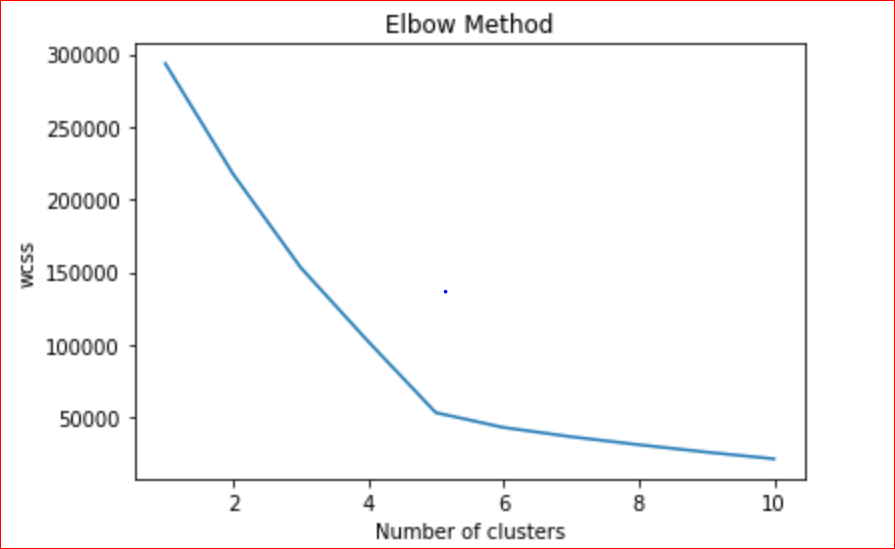
* wage per hour
* weeks worked in a year
* Age
* capital gain
* capital loss

Before implementing K mean, we have to find out the **optimal number of clusters** that would be best suitable for the clustering. This can be done using elbow method.

1. **Number Of Cluster (Elbow Method):**

When using k-means, we need to determine the right number of clusters. This can be done more or less accurately by iterating through different values for the number of groups and compare an amount called **the within-cluster sum of square distances** for each iteration. We normally stop when the improvement in this value starts decreasing at a lower rate.

I used sklearn in order to perform clustering on the data.



Based on the elbow plot, we could choose between 4 and 6 clusters.I choose optimal number of cluster as 5.

1. **Dimension reduction (PCA):**  Principal Component Analysis to transform our multi-dimensional dataset into a 2 dimensional dataset. As there are more than 2 features I am using to make the cluster, used PCA to reduce the dimension, so that the we can easily see the cluster in 2d plane.
2. **Model fitting:**

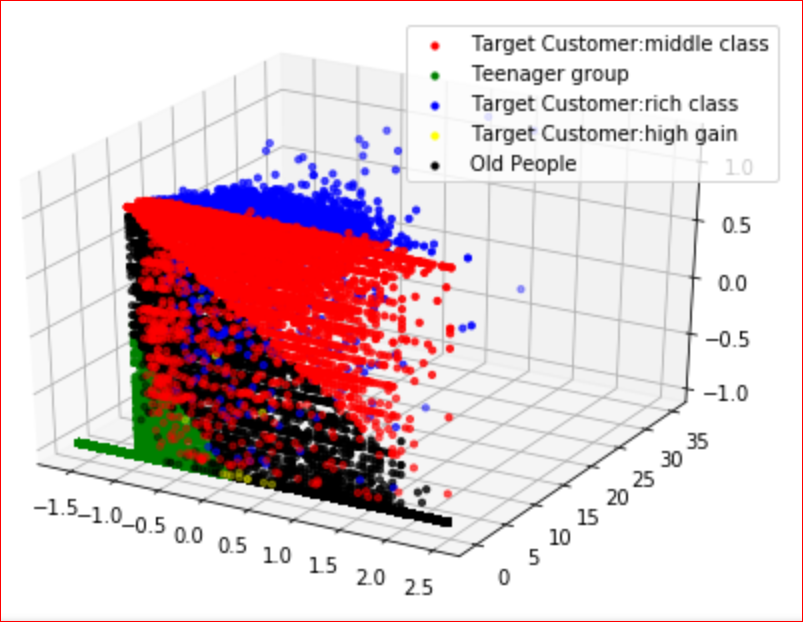
 I have used the k-means algorithm from scikit-learn to fit the model. The number of cluster used is based on optimal cluster found in elbow method.

1. **Test Prediction:**

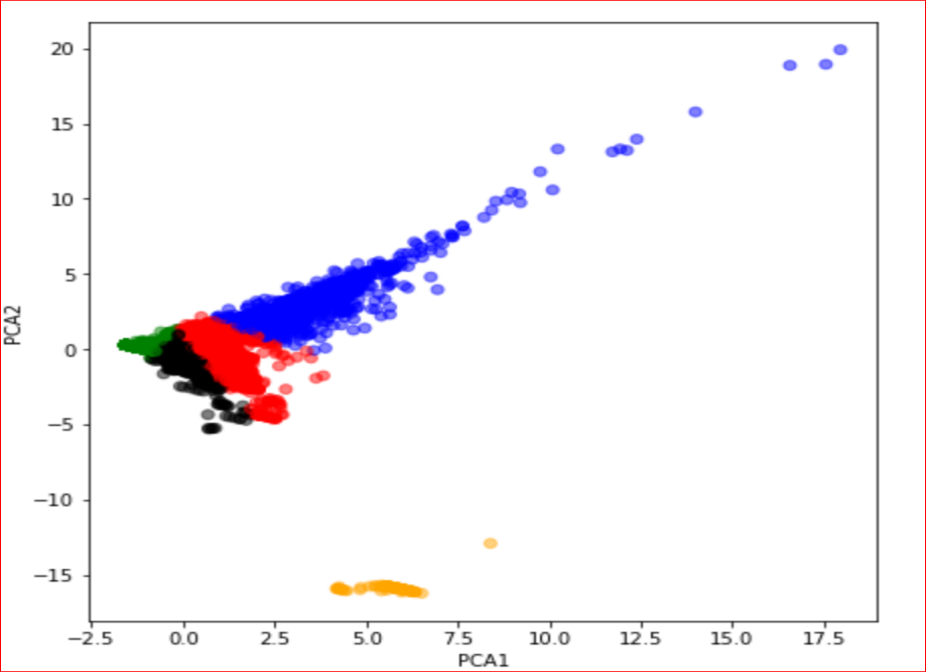
After training and finding respective cluster, I test the position of centroid on the new test data set .

1. **Visualization:**

The 3d Visualization of cluster.



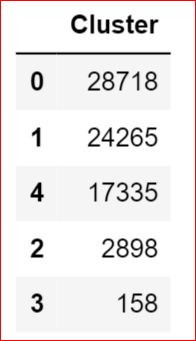
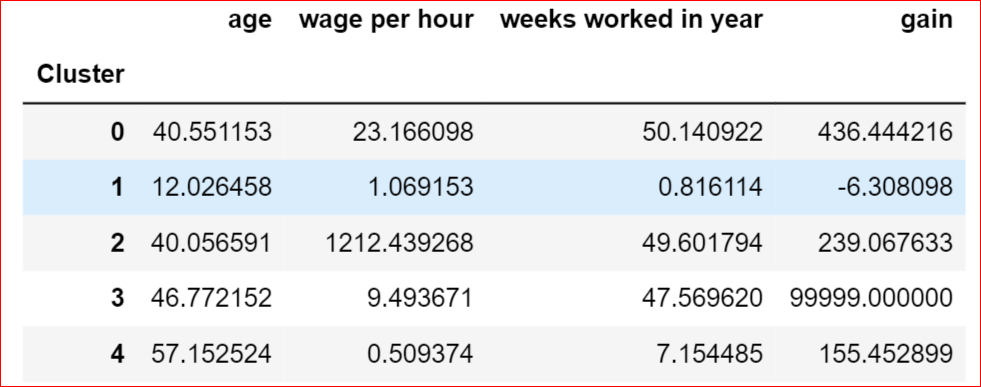
2d visualization obtained after PCA.



1. **Cluster Analysis:**

I break out cluster and compare it to the remaining customer clusters to see how one group of people are different from the other group.

**Summary Statistic:**



**Cluster 1** (**Middle-class Target Profile**): Medium wage per hour (23) but mostly they have worked for the whole year (worked for 50 weeks in a year) and net gain is also moderate. The average age of this group of people is around 40. This group can be called as Medium Profile group and these can be the target group for retail marketing . Number of people in this group is **28718**.

**Cluster 2 (Teenager Profile):** This group has Low wage per hour and very less week worked in a year with average age 12. The gain is -ve. All these things imply that most people in this group might be students . This group might not be a targeted group for retail marketing.

**Cluster 3(high-class target profile):** These group of people has a high wage per hour, have worked for more weeks in a year and moderate capital gain. This group can be the target group as their wage per hour is high, average age lies around 40 and these people have worked throughout the year, this group can also be one of the targeted customer for Walmart. These people might be wealthy, So in regards to the market, these people should be targeted to market for an exclusive product and not the normal product. Number of customer here is **2898**

**Cluster 4(Target profile):** These group of people has less wage, high work per year, High capital gain. This group can be the target group as the wage per hour is less but the gain is high. Age group lies around 46 and these people have worked throughout the year. As they have high gain Walmart might find promotion/offer related to that. Very fewer people are in this group, around **158**

**Cluster 5(Old profile):** People this group has an average age of 60, less wage and less week worked per year with low capital gain. This group might not be the target group due to their age or wage.

As per my research Walmart’s target customer are middle age people.

#### **How this will help?**

For new promotional offer, instead of calling each customer, first Walmart will focus on "Target" group of customer i:e customers in red, yellow and blue zone.

**Different type of promotional offer will suit for different type customer group . Based on each cluster of customer Walmart can choose different strategy of promotion.**

1. **Model Evaluation:**

### **Classification Accuracy:** The correct predictions to the total number of input samples. The accuracy is found to be 86%.

### **Confusion Matrix:** Confusion matrix allows to look at the particular misclassified examples.

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Predicted* | | |
| *Actual* |  | *0 (<50k)* | *1 (>50k)* |
| *0(<50k)* | *TN (****38227****)* | *FP (****7779****)* |
| *1(>50k)* | *FN (****5073****)* | *TP(****40471****)* |

**True Positives :** The cases in which predicted is > 50k and the actual output was also >50k.

**True Negatives :** The cases in which we predicted <50k and the actual output was >50k.

**False Positives :** The cases in which we predicted >50k and the actual output was <50k.

**False Negatives :** The cases in which we predicted <50k and the actual output was also <50k.

**Observation:**

The result is telling us that we have 38227+40471 correct predictions and 5073+7779 incorrect predictions.

4.3 **Classification Report:**

|  |
| --- |
| precision recall f1-score support  <50 0.88 0.83 0.86 46006  >50 0.84 0.89 0.86 45544  avg / total 0.86 0.86 0.86 91550 |

Precision: Precision for both classes I:e <50k and >50k are around 0.88 and 0.84, which is a good

score.

Recall: Fraction of positives to that were correctly identified.

F1-Score: The f1-score gives you the mean of precision and recall. The scores corresponding

to every class will tell the accuracy of the classifier in classifying the data points in that

particular class compared to all other classes.The F score reaches its best value at 1 and worst

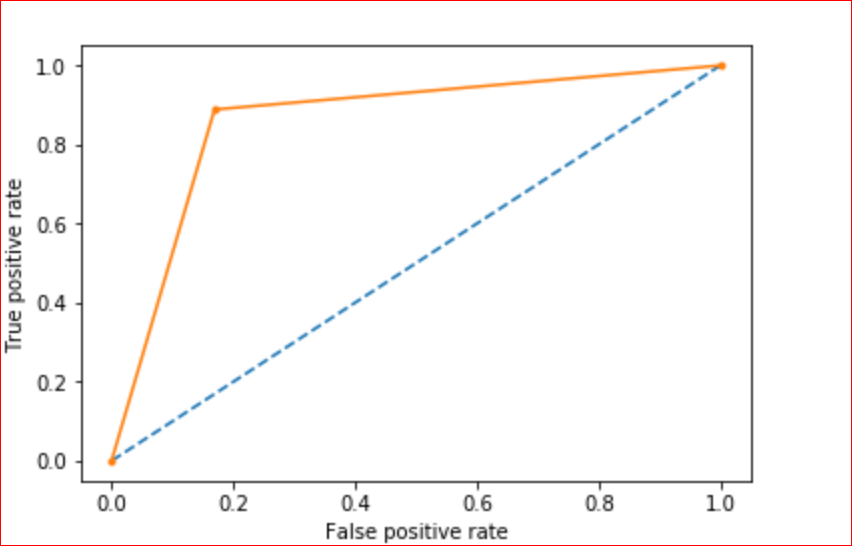
score at 0. For our data set the f1- score is 0.86 which can be consider as a good score.

4.4 **ROC Curves and AUC :**

Area under ROC Curve (or AUC for short) is a performance metric for binary classification problems.

The AUC represents a model’s ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random.So any value close to 1 can be consider as good model.

The AUC for our model is 0.859.



**Programming Languages:** Python

**References:**

*<https://www.marketing91.com/marketing-strategy-walmart/>*

*<https://www.marketwatch.com/story/walmart-earnings-e-commerce-investments-could-offset-earnings-growth-2018-08-13>*

*<https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/>*

*<http://www.kimberlycoffey.com/blog/2016/8/k-means-clustering-for-customer-segmentation>*

*<https://en.wikipedia.org/wiki/K-means_clustering>*

*<https://hackernoon.com/introduction-to-machine-learning-algorithms-logistic-regression-cbdd82d81a36>*

*[https://xgboost.readthedocs.io/en/latest/python/python\_api.html#xgboost.XGBClassifier.fit](https://xgboost.readthedocs.io/en/latest/python/python_api.html" \l "xgboost.XGBClassifier.fit" \t "http://localhost:8888/notebooks/MLExercise/_blank)*