**CENSUS INCOME**

By Prachi Parmar

**PROBLEM DEFINATION**

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of reasonably clean records was extracted using the following conditions:

((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). The prediction task is to determine whether a person makes over$50K a year.

**DATA INFORMATION**

**FEATURE INFORMATION:**

***Age****: Ranges from 27 to 90*

***WorkClass****: Categorical Feature (Private, Self-emp-not-inc, local-gov, State-gov, self-emp-inc, federal-gov, without-pay, never-worked)*

***Fnlwgt****: Ranges from 12,285 to 14,84,705*

***Education****: Categorical Feature (Bachelors, HS-grad, 11th, Masters, 9th, Some-college, Assoc-acdm, Assoc-voc, 7th-8th, Doctorate, Prof-school, 5th-6th, 10th, 1st-4th, Preschool, 12th)*

***Education\_num****: Ranges from 1 to 16*

***Martial\_status*** *: Categorical Feature (Married-civ-spouse, Divorced, Married-spouse-absent, Never-married, Separated, Married-AF-spouse, Widowed)*

***Occupation:*** *Categorical feature* (Exec-managerial, Handlers-cleaners, Prof-specialty, Other-service, Adm-clerical, Sales, Craft-repair, Transport-moving, Farming-fishing, Machine-op-inspct, Tech-support, Protective-serv, Armed-Forces, Priv-house-serv)

***Relationship:*** *Categorical Feature (Husband, Not-in-family, own-child, unmarried, wife, other-relative)*

***Race:*** *Categorical Feature (White, Black, Asian-Pac-Islander, Amer-Indian-eskimo, Others)*

***Sex :*** *Male/ Female*

***Capital\_gain:*** *Ranges from 0 to 99,999*

***Capital\_loss:*** *Ranges 0 to 4356*

***Hours\_per\_week :*** *Ranges from 1 to 99*

***Native\_country :*** *Categorical Feature containing countries.*

**TARGET VARIABLE:**

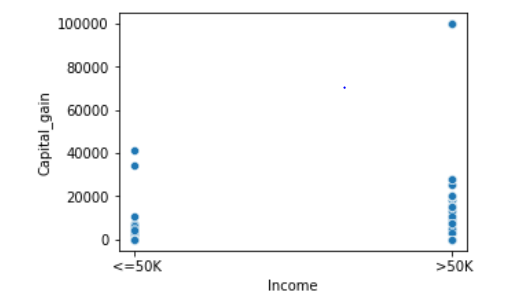
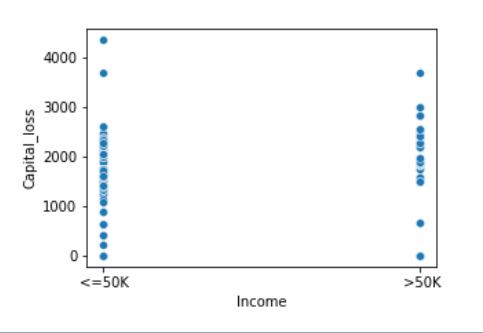
***Income****: <=50K / >50K*

**DATA ANALYSIS**

* 5 continuous predictors, which includes Age, Fnlwgt, Capital gain, Capital Loss and Hours\_per\_week.
* 9 categorical predictors, including Native country, sex, race, relationship, occupation, marital status, work class, education, education num.
* Age Feature has extreme values as its min value is 17 and .75 percentile or 75% of the data points of age is below 48 and difference between .75 and max value is very high, suggesting high variance in data and causing skewness in data.
* Capital Gain and Capital Loss has 75% of Data points as value 0. Capital gain has 25% of data between 0 and 99,999. Capital loss has 25% of data points between 0 and 4356.
* Work Class, Occupation and Native Country has missing values with ‘?’ tags. It can be treated with mode values of the Column as missing values are present in ordinal feature.
* Target Variable has 2 classes making it binary classification.
* Data Types of the features are int64 for continuous features and object type for all the ordinal data types.

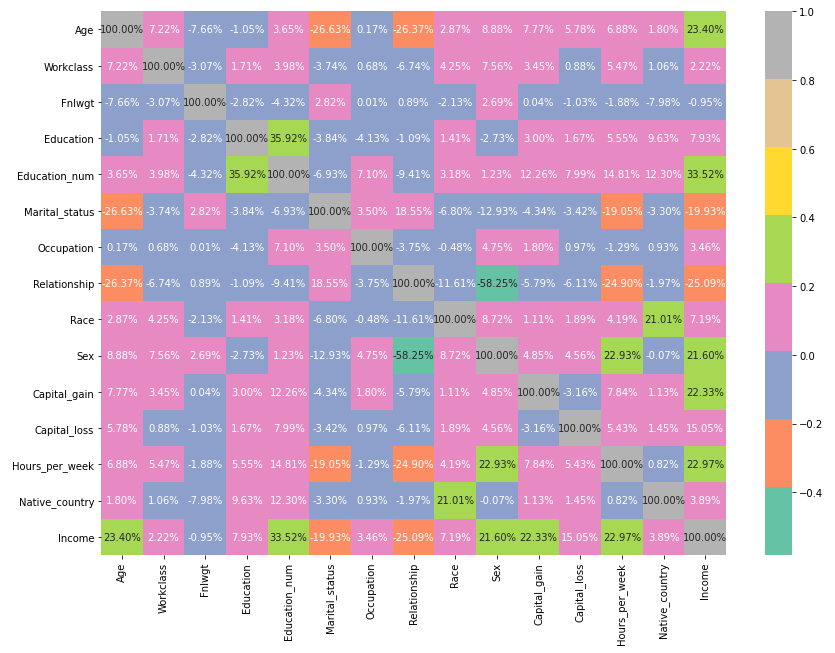
**EDA REMARKS**

* Target label have imbalanced classes with <=50K having 75% of data points and >50K with 25% of data points.
* Applying under sampling to balance the target classes.
* 69% of the population data works in private company, 11% of population are self-employed including with and without income, 13% of population works in government based work environment. Self-employment with income part of population have high probability (around > 80%) to have income >50K. Federal Government work class also has more than 70% probability than other categories to have >50K income category. Other work classes have same probability than other categories to belong in >50K or <-50K category. Never-worked, without pay always has <=50K, we can assume this category has high probability(100%) than other categories to belong to <=50K .
* Highest count of the population are with a HS-Graduate degree with 32% of the population. Population part with Bachelors degree have high probability to belong in >50K category. Most of the population with school degree have high probability to belong in <=50K. Population with some college and HS-Grad have high probability to belong in <=50K. Population with Master’s and Doctorate degree have probability to belong in >50K.
* Highest count of population are married with civil spouse with 45% of population. 32% of population are unmarried. Widowed, Separated, Never-Married, Divorced have high probability to belong in <=50K. Married-cis-spouse has high probability to belong in >50K.
* Population with occupation Exec-managerial, prof-speciality have high probability to belong in >50K. Adm-clerical, other services, transport moving have high probability towards <=50K.
* Highest count in the population has relationship as Husband, 10% are unmarried and 15% are own-child. Most of the population with relationship as wife or husband have higher probability (75% +) than other categories to belong in >50K. Own Child, Other-relative unmarried has higher Probability (85% +) than other categories to belong in <=50K. Not in family has 75%+ probability than other categories for <=50K.
* Most of the population people are White race and have 55:45 probability to belong in >50: <=50K. Population with Black Race have higher probability to belong in <=50K category.
* Male population has higher probability than female population to belong in >50K and Female population has higher probability than Male population to belong in <=50K.
* Native Country has equal probability to belong in both of the income categories.
* More than 90% of the population is From USA and other 10% belong is different countries.
* Education and education number are related. For lower education number, more population count in <=50K and for higher education number, high population count for >50K.
* Mean Age of >50K is higher than <=50K.
* Mean Capital gain is very high for >50K than <=50K. Mean Capital gain for >50K is 4k and that of <=50K is 150 approximately.
* Mean Capital loss is very high for >50K than <=50K. Mean Capital loss for >50K is 53 and that of <=50K is 158 approximately.
* Mean value for Hours\_per\_week for >50K is higher than <=50K. Mean age for <=50K is 45 and that of >50K is 38.
* Mean value for Fnlwgt is same for both target classes. Fnlwgt has less effect on target variable.

* From the above graph, we can say capital gain greater than 1900 has higher probability to be in >50K income bracket. For capital loss lesser than 1500 has higher probability to belong in <=50K.

**Correlation Table**



* From the correlation table, multicollinearity can be seen between predictors.
* Multicollinearity between Age and Relationship/ Marital Status.
* Multicollinearity between Education and Education number.
* Multicollinearity between Marital Status and hours per week/ Age.
* Multicollinearity between Relationship and Sex/ hours per week/ Age.
* Multicollinearity between Native Country and Gain
* Moderate Correlation about 20-35% with Target Variable Income is highest with Hours per week, capital gain, capital loss, Relationship, Marital Status, Education Number and Age.
* Distribution of Continuous were left skewed structure for capital gain, Fnlwgt and capital loss. Hours per week and age has more gaussian type structure.
* Boxplots helped in determining whether a feature has outliers or not. Capital loss, Capital Gain and Fnlwgt has outliers. Outliers in categorical data are not treated as outliers.

**PRE-PROCESSING PIPELINE**

Pre-Processing the data involves:

* Outlier Detection
* Outlier Treatment
* Feature Selection via correlation and VIF (variance influence factor)
* Standardization/ Scaling of Predictors.

**Outlier Detection:**

Code snippet:

q1= x.quantile(.25)

q3 = x.quantile(.75)

iqr = q3-q1

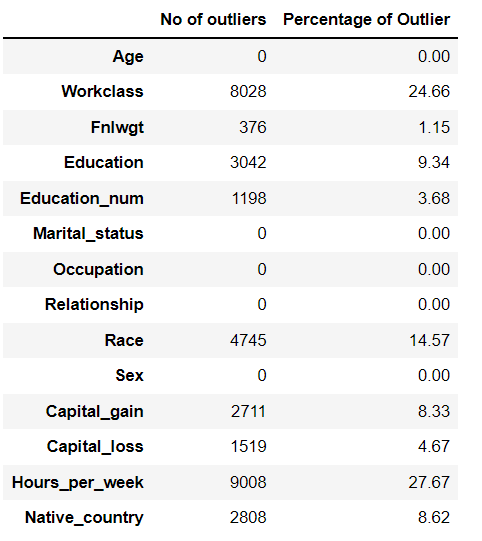
count=((x.iloc[:] < ( q1 -1.5\*iqr ))|( x.iloc[:] > (q3+1.5\*iqr ))).sum(axis=0)

count= (( x.iloc[:] < (q1-1.5\*iqr) )| (x.iloc [:] > ( q3+1.5 \* iqr ))).sum(axis=0)

outlier\_table= pd.DataFrame(count,index=x.columns,columns=['No of outliers'])

outlier\_table['PercentageofOutlier'] = round(outlier\_table['No of outliers']\*100/len(x),2)

>OUTPUT

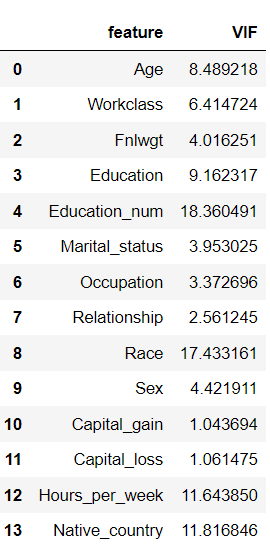


* As the number of outliers in continuous predictors are more than 30%. We can treat the outliers with transformation rather than removing the outliers.

**Transforming the data**

* Transforming the data with Power Transformer to scale the data as Power-Transformer scales the data present with outliers and also reduces the skewness in the gaussian like structure feature.

**Feature Selection**



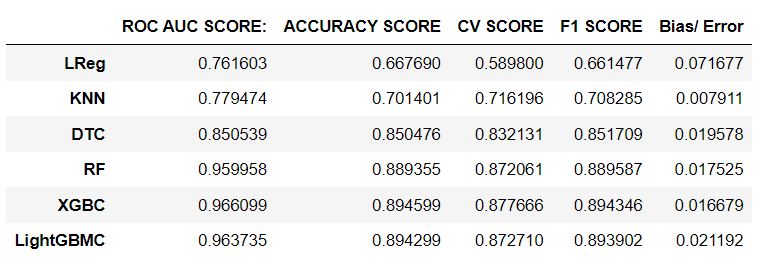
* Keeping the VIF threshold as 10, we see multicollinearity presence in some features.
* Features dropped are : ['Education\_num','Race'] and After removal of Education number and Race, we have all feature VIF <10.
* Scaling of Data is done with Power-Transformer.
* Last step in pre-processing the data is under-sampling / over-sampling the data in presence of imbalanced target classes.
* Over-Sampling the data with SMOTE:

x\_over, y\_over = SMOTE().fit\_resample(x1,y)

**BUILDING MACHINE LEARNING MODELS**

* We have trained the dataset with various models like Logistic Regression, K Nearest Neighbours, Decision Tree Classifier, Random Forest Classifier, XG Boost Classifier, LightGBoost Classifier.
* Evaluation metric for this problem is F1-Score and Accuracy Score due to imbalanced target classes. Due to Imbalanced target Classes, need to check if the results are not biased towards majority target class.
* F1 score is harmonic mean of precision score and recall score, and classification report can give f1 scores, precision, recall scores for each target class to check biasness in the minority class.
* Training Score will be calculated with Cross Validation method where train set is further divided into n iterations set(n-1 part for train and 1 part for test ) and mean of each iteration is taken as the final training score where in every iteration train and test set is different .

**SCORE TABLE OF ALL MACHINE LEARNING MODELS**



* As our data contains more categorical data than continuous data, Logistic Regression model does not give accurate as other models. Logistic regression tend to make linear relationship with features and in case of categorical predictors, logistic regression works not up to mark.
* Bagging model, random forest classifier model gives f1-score up to 88.9% and .95% roc auc score.
* Boosting models XGB Classifier and Light GBM Classifier gives up to .965% roc auc score up to 89.4 f1 score.
* Error term is difference between testing score, which in our case is f1-score and Training score, that is cross validation score.
* We choose a model with least error, as high error term (difference in testing and training scores) would suggest overfitting/ underfitting of the model.
* Choosing Random Forest model, XGB and Light GB Model for hyper parameter tuning.

**Hyper-Parameter Tuning**

After Tuning Random Forest, XGB and Light GB, Light GB gave best results with least error.

* LIGHT GB HYPER PARAMETERS

params\_lgbr = {

'num\_leaves': [70,80,50,40,20],

'max\_depth': [5,7,10,12,15],

'bagging\_freq': [5,10,15,7,12],

"bagging\_fraction": [0.75,.6,.5,.8,.85,.7],

'learning\_rate': [.01,.02,.1,.5]

}

* LightGBM Classifier hyper tuning model fitting code snippet:

LightGBR = LGBMClassifier()

grid\_l= RandomizedSearchCV( LightGBR, params\_lgbr, cv=5, scoring='f1', n\_iter = 30, verbose=2, n\_jobs=5)

* 50 iterations of randomized search with the parameters above, we get the best parameters for the model that gives highest accuracy score.
* Best Parameters are:

{'subsample': 0.8,

'n\_estimators': 200,

'min\_child\_weight': 2,

'max\_depth': 15,

'learning\_rate': 0.2,

'gamma': 2}

**FINAL TUNED MODEL EVALUATION: LIGHT GB MODEL**

Accuracy Score : 0.906

ROC AUC Score : 0.973

Precision Score : 0.906

Recall Score : 0.907

F1 Score : 0.9067980001492426

**(TESTING SCORE)**

Cross Validation Score : 0.8928594784218997

**(TRAINING SCORE)**

**CLASSIFICATION REPORT:**

precision recall f1-score support

0 0.91 0.91 0.91 6654

1 0.91 0.91 0.91 6695

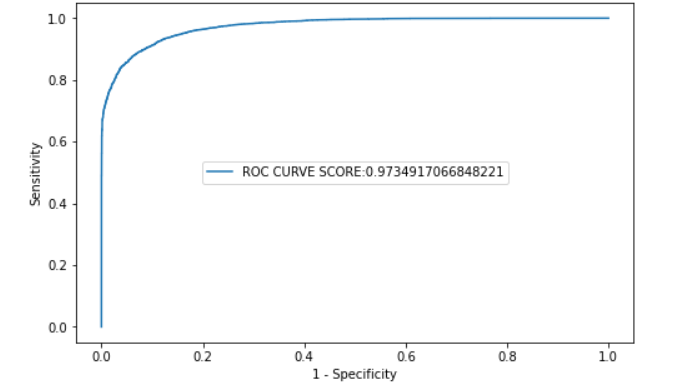
accuracy 0.91 13349

macro avg 0.91 0.91 0.91 13349

weighted avg 0.91 0.91 0.91 13349

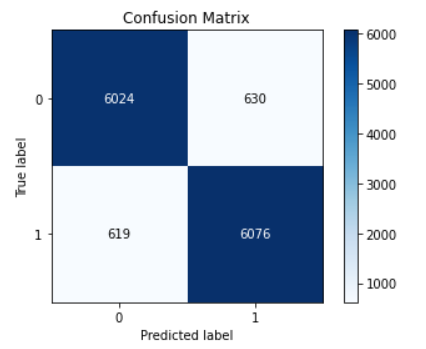
* According to the Classification report, precision, recall, f1 scores for both target labels is same.
* No biased results for minority class, models works on more than 90% accuracy for both target class (majority and minority target class)

**ROC AUC CURVE:**



* ROC AUC Score is the probability of predicting the positive label rate, that is 97.3% for our chosen model.
* 97.3% of the times, correctly predicted positive label.

**CONFUSION MATRIX**



* Confusion matrix is matrix of correctly predicted count of each label and falsely predicted count for each label. Correctly predicted labels are True Positive and True Negatives. Falsely predicted Labels are False Positive and False Negative.
* From the confusion matrix, count of False Positive and False Negative are lowest.
* **SAVING AND LOADING THE MODEL WITH JOBLIB**
* Saving the tuned model with joblib.dump and can load the model for future prediction with joblib.load
* **import** **joblib** **as** **jb**
* jb.dump(model,'census\_lgbc.pk1')
* **LOADING THE MODEL FOR FUTURE PREDICTIONS fOR CENSUS INCOME**
* jb.load('census\_lgbc.pk1')
* OUTPUT
* > LGBMClassifier(bagging\_fraction=0.75, bagging\_freq=15, max\_depth=24,
* num\_leaves=100)

**OVERFITTING COMMENTS:**

When test score and train score difference is very high and test score > train score , it is overfitting.

As out model has good accuracy score of 91% and test score of 90%, we can rule out overfitting as difference between the scores are very less.

**CONCLUDING REMARKS**

* Achieved a test score (f1 score) of 91% after hyper tuning boosting technique LightGBM Classifier.
* Achieved a train score (cross validation score with 5 iterations) for the model.
* Achieved a Accuracy score of 90.5% for the model.
* Achieved a roc auc score of 97.3% for the model.
* No biased results. Test score same and high for both target classes.
* As error term between test score and train score is very low, we rule out overfitting or underfitting.
* Census Income is highly correlated with Education, Age, Hours\_per\_week and Sex.
* From visualization, Capital gain and Capital loss have major differences in the mean values for both the census income bracket.