CAPSTONE PROJECT ON
ANALYSIS OF CENSUS INCOME DATASET

Submitted in Partial Fulfillment of requirements for the Award of certificate of
Post Graduate Program in Data science and Engineering
(PGPDSE July 2019 to December 2019)

Project Report
Submitted to

GREAT LAKES
INSTITUTE OF MANAGEMENT
Global Mindset - Indian Roots

SUBMITTED BY:

ABHEESH NAIR

AYUSHI JAIN

UNDER THE GUIDANCE OF:

MR. JAYVEER NANDA

(Mentor)

Abstract: Developed a stable and optimized model to predict the income of the individual whether a person makes over $50k a year using logistic regression along with ensemble machine learning techniques.

Tools and Techniques: Python, Tableau, Logistic Regression, Random Forest, Gradient Boosting, Decision Trees, KNN

ACKNOWLEDGEMENT

We take this opportunity to express our profound gratitude and deep regards to our mentor **Mr. Jayveer Nanda** for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. The help and guidance given by him time to time shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to the faculty and management office of Great Lakes Institute of Management for their support, valuable information and guidance, which helped us in completing this task through various stages. We are grateful for their co-operation during the period of our project.

Lastly, we thank almighty, our family and friends for their constant encouragement without which this course would not be possible.

Abheesh Nair

Ayushi Jain

Contents

Dataset Introduction:

In our dataset there are thirty-two thousand records and a binomial label indicating income of less or greater than fifty thousand US dollars, referred to as <=50K or >50K in this report. It is an imbalance data with 76% of the records in the dataset with a class label of <=50K.

There are fourteen attributes consisting of seven polynomials, one binomial and six continuous attributes. The nominal employment class attribute describes the type of employer such as self-employed or federal and occupation describes the employment type such as farming or managerial. The education attribute contains the highest level of education attained such as high school graduate or doctorate. The relationship attribute has categories such as unmarried or husband and the marital status attribute has categories such as married or separated. The final nominal attributes are country of residence, gender and race. The continuous attributes are age, hours worked per week, education number (which is a numerical representation of the nominal education attribute), capital gain and loss and a survey/final weight attribute which a demographic score is assigned to an individual based on information such as area of residence and type of employment.

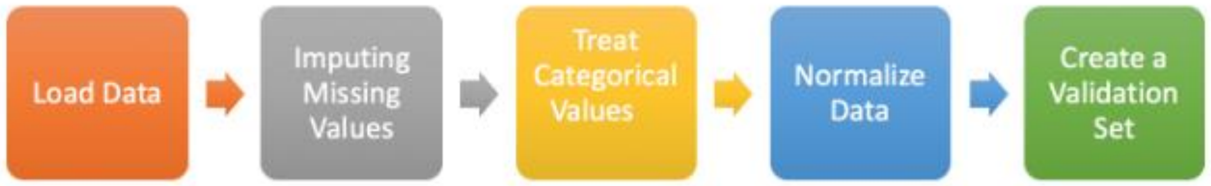
Objective:

The purpose of this project is to build a model to predict income of countries in a given data set with features such as education, hours-per-week, marital status etc. so that the government can modify their policy and get a meaningfully picture about the population.

1. To analyze if the person gets more income if he works for longer hours.
2. To analyze whether the person is under-paid when all the factors (Education, work-class etc.) are favorable.
3. To understand and analyze the capital gain and capital loss mechanism on the income level.
4. To analyze whether age plays a significant role in deciding the income level of an individual.
5. To predict whether the individual has more than $50k income or not.

How We Progressed:

We started with the basic work which is familiarizing ourselves with the dataset attributes and what is the significant of each attribute. We then addressed the missing values or null values which is in the form of '?' and performed exploratory data analysis which is descriptive statistics and visualization. Then after this we did some inferential statistics which is hypothesis testing and Even though relation between variables with target are usually seen in graphs, this is not enough to prove that these columns are not just by chance. We must share proof of existence of their relation with the help of statistics. To find the statistical significance we applied ANOVA test for categorical vs numerical variables and chi square test for categorical vs categorical variable and making contingency table and after all this we implemented several machine learning techniques, including Logistic Regression, Decision Tree Classifier, bagging methods including, Random Forest Classifier and Boosting methods such as Gradient Boosting and Adaboost. We evaluated our models by calculating the Accuracy, AUC-ROC curves and used confusion matrix to find f1score, precision and recall to see which model performs the best.



Load Data → Imputing Missing Values → Treat Categorical Values → Normalize Data → Create a Validation Set

Data Cleaning:

Data Cleaning is the first step for any dataset as we cannot build our model on uncleansed data and we also cannot rely on the accuracy score of that model. Thus, it is mandatory to clean the dataset first and then start working on it.

As in machine learning some people say “Garbage IN Garbage OUT “.

We see some unknown/special characters like ‘?’ in the dataset which are incorrectly entered in our data. So, we replace these characters and here we start with the cleaning process of the dataset. We use replace command and change it to np.nan and null values which can be imputed later by finding some pattern among variables having non null values and the column with null values.

We see null values are present in columns work-class, occupation and native country. This needs to be filled.

Missing Value Imputation and Category Reduction/Generalization:

In the Dataset people mostly belong to Private sector job and so we fill our missing value in work class column with its mode and we are reducing the number of categories by generalizing the ‘workclass’ according to the same class like all the gov. in one department.

We find null values in occupation column for which we must find some pattern and fill the data.

We are reducing the number of categories by generalizing the occupation according to the designation of workers by collar colors.

- Blue Collar
- Pink Collar
- White Collar
- Gold Collar

We actually get good observations about occupation on basis of education people have done. Like people with only High school education are either blue collars or pink collars.

And people with graduation or post-graduation are having White collar and Gold Collar jobs respectively. So, we fill with the mode values accordingly.

Similarly, for education column we reduce the columns by generalizing the information in it. We provide on category to people educated from 1st to 6th class and so on for all the education levels. The idea behind this is to fill the null values in occupation column using the mode metric of education column for each category.

In the same way we will generalize and reduce the categories in the marital status column.

We are converting this column to 4 categories named as Not Married, Married, Separated, Widowed.

In column native country null values are considered as ‘Others’ as the mode of US is very high so we changed all other country with others which can be seen in Fig 1.1 and Fig 1.2 below.

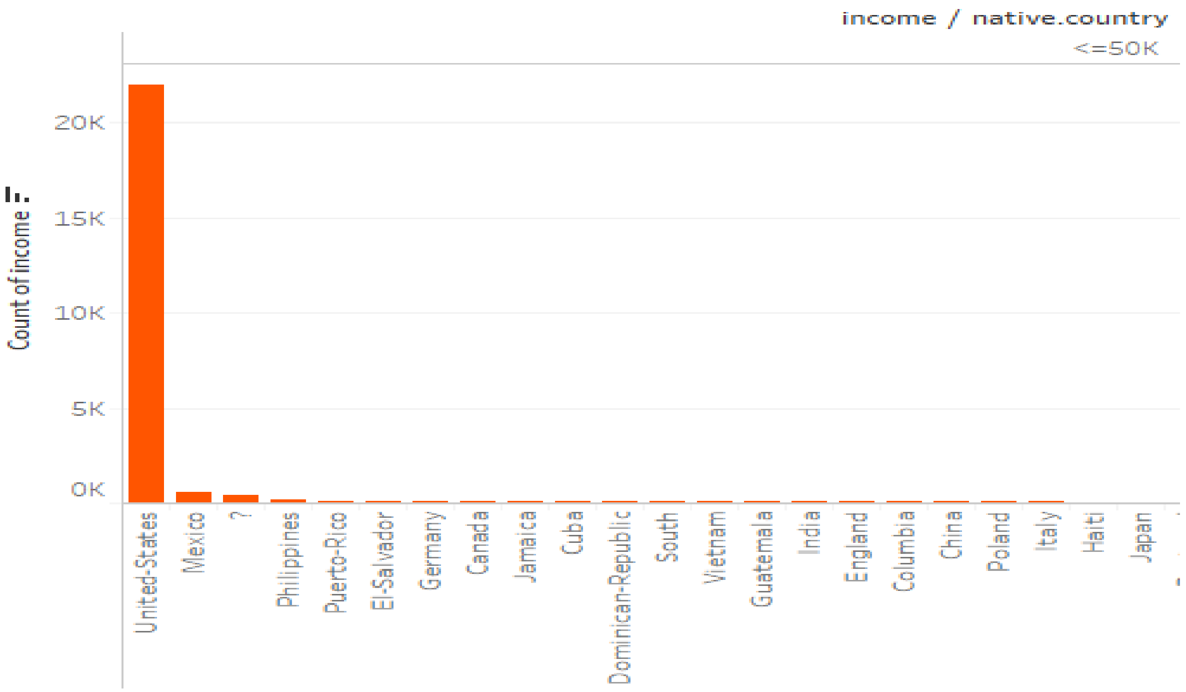


Fig-1.1: Depicts the countries with Income count <= 50 K

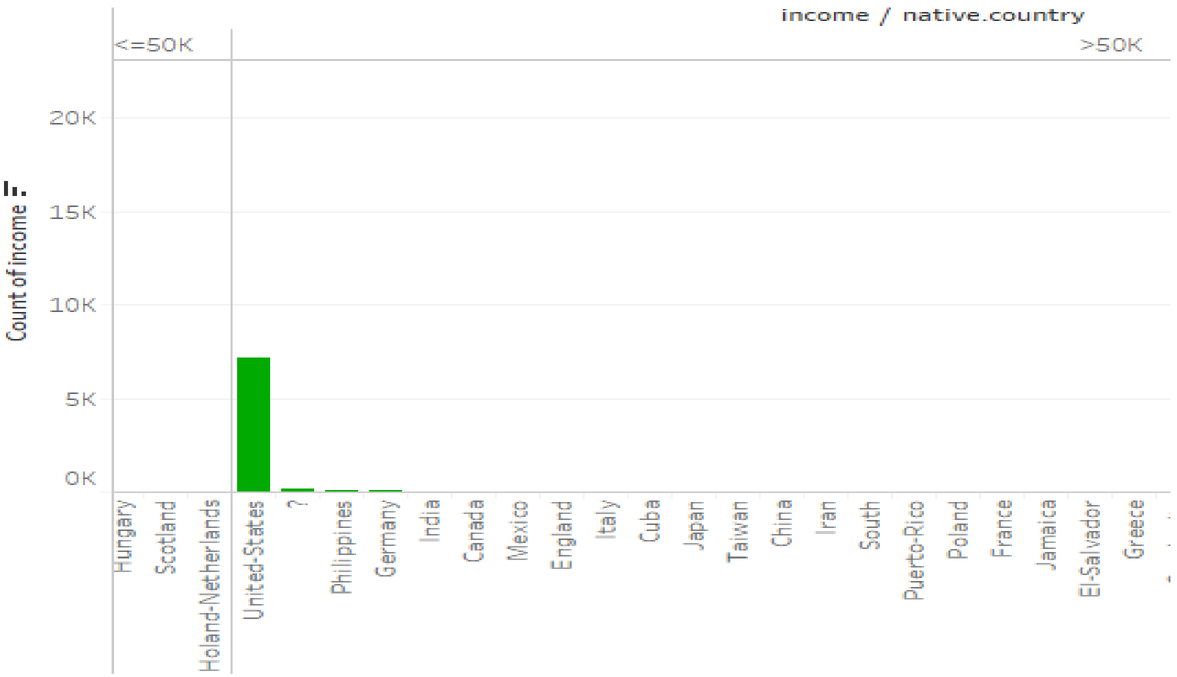


Fig-1.2: Depicts the countries with Income Count > 50 K

Descriptive Statistics:

Univariate Analysis

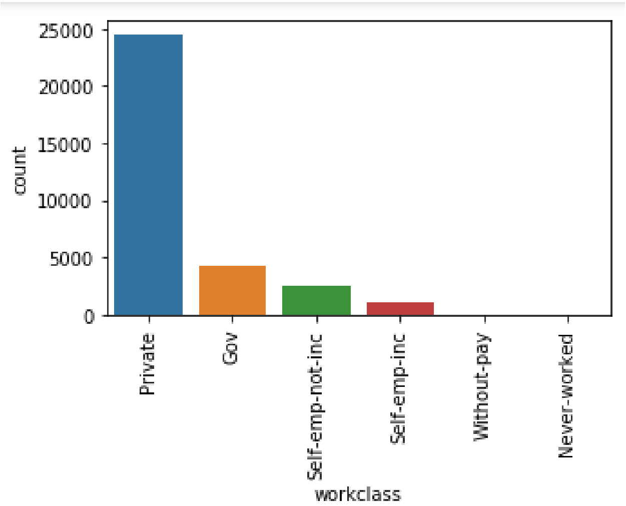


Fig-1.3 : Depicts the relationship between the Workclass and the Count of Individuals

The above illustration exhibits the relation that a majority of Individuals belong to the Private workclass as compared to other workclasses such as Government, self-employed etc

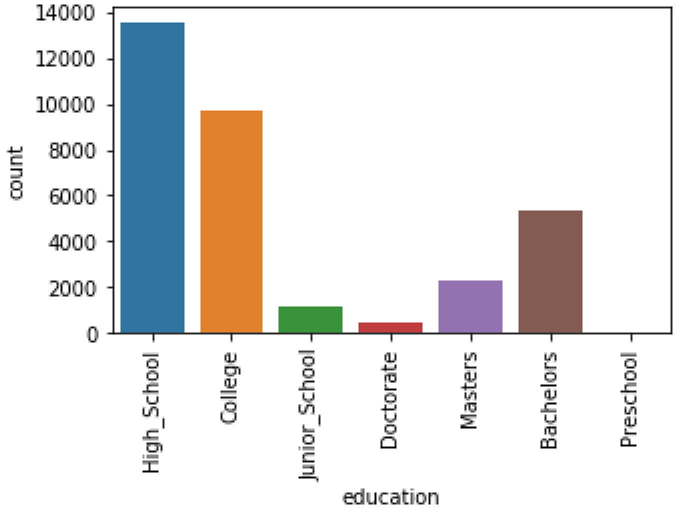


Fig-1.4 : Depicts the relationship between Count of individuals and Education

The above illustration exhibits that a majority of individuals are literate. It is also evident that a majority of them started working post their High-school education.

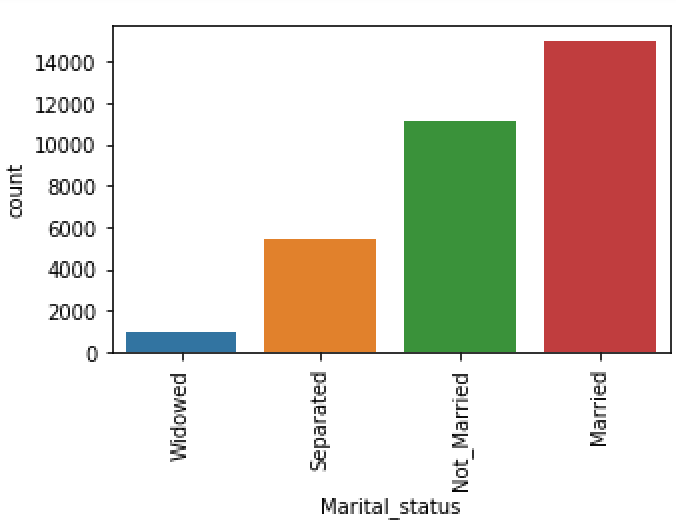


Fig- 1.5: Depicts the relationship between the count of individuals and their marital status

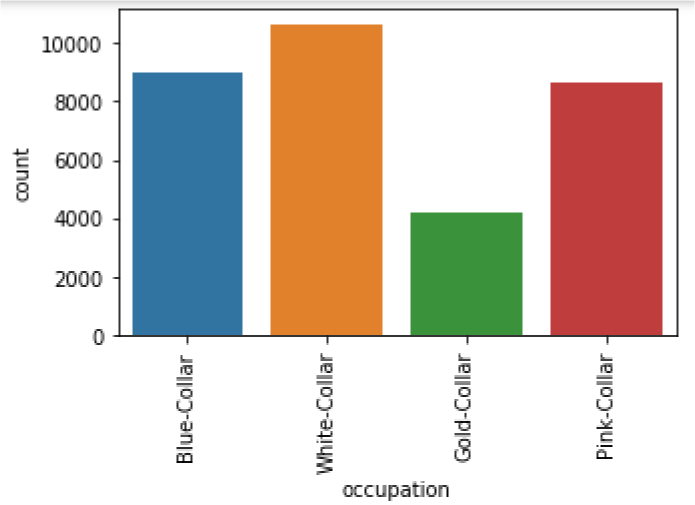


Fig – 1.6: Depicts the relationship between the count of individuals and their Occupation

The above illustration exhibits nearly equal proportions when the different classes of occupation are being taken into consideration. This signifies almost equal importance of each occupation category in the economy

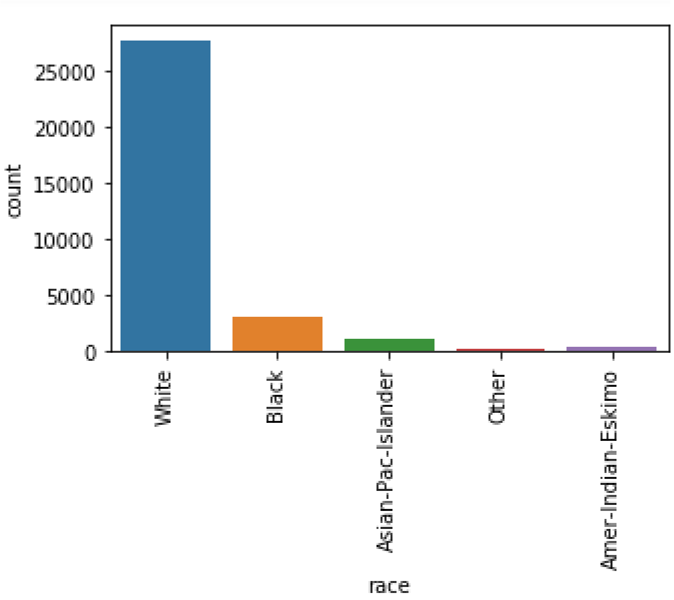


Fig- 1.7: Depicts the relation between the Count of individuals and their Race.

The above illustration exhibits the ratio of individuals of different races in the given dataset. It is observed that a majority of the population belongs to the race "White" followed by "Black" and others.

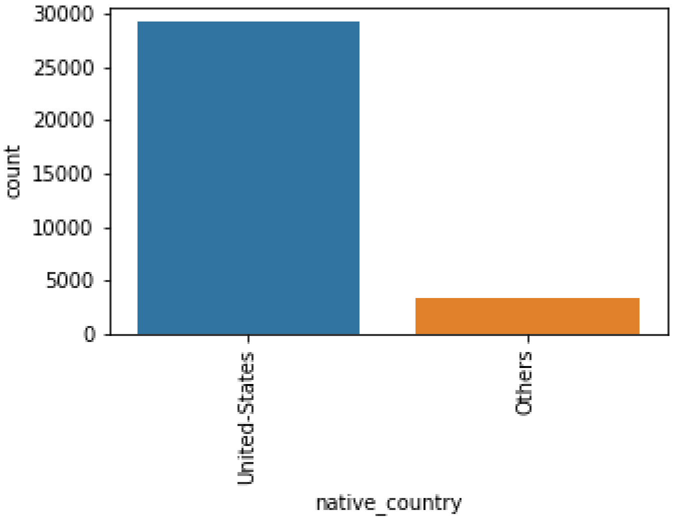
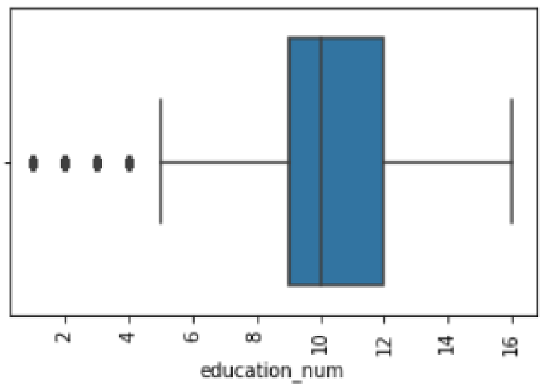
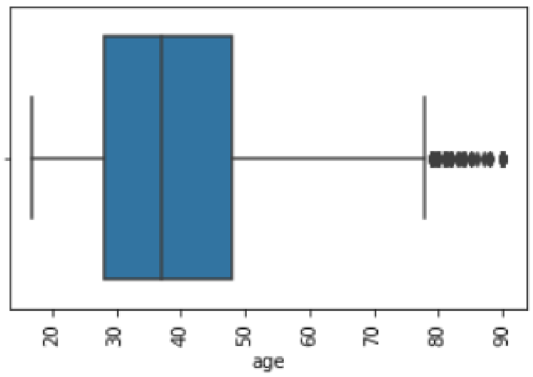
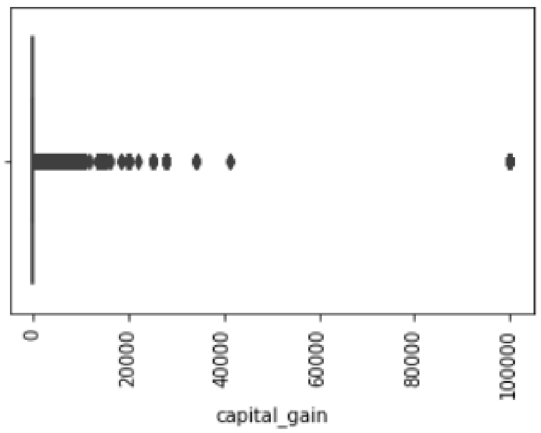


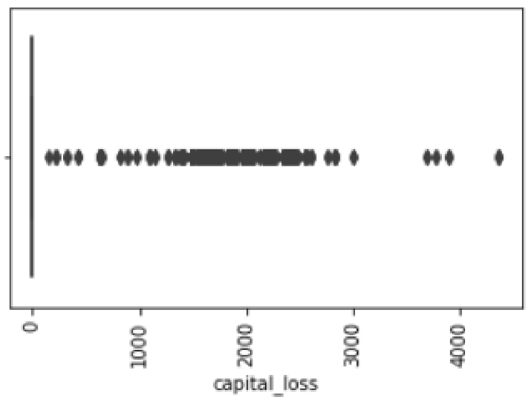
Fig-1.8: Depicts the relationship between the count of individuals and native country

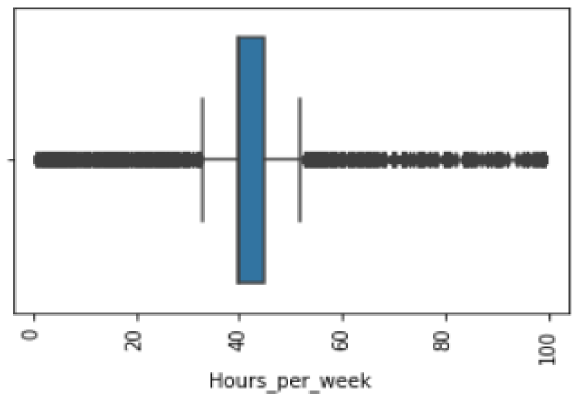
The above illustration exhibits unanimous majority of individuals belonging to the United States of America when the category "native_country" is being taken into consideration.

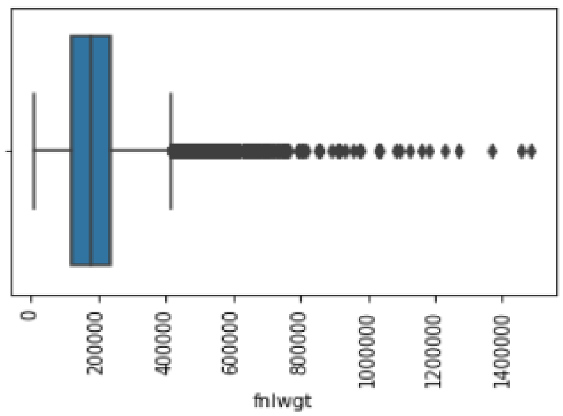












Analysis of Boxplots

The above illustrations exhibit the distribution of categories ‘education.num’ , ‘age’ , ‘capital_gain’ ,’capital_loss’ ,’Hours_per_week” and ‘fnlwgt’ in the form of boxplots.

Bivariate Analysis

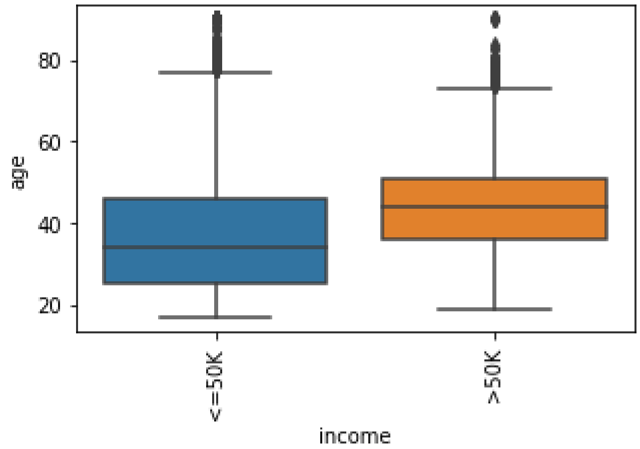


Fig -1.9: Depicts the relationship between Age and Income

It is observed that younger people tend to earn less income when compared to older people.

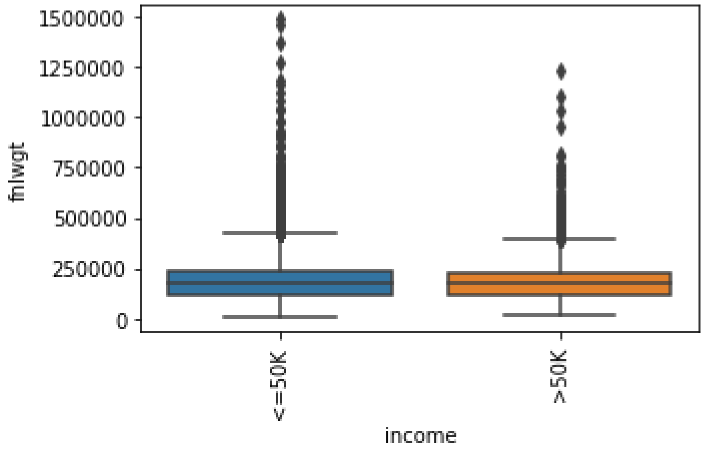


Fig- 1.10 : Depicts the Relationship between Final weight and Income

The boxplot exhibits no significant impact of Final weight on Income. We shall perform the ANOVA test for further insights.

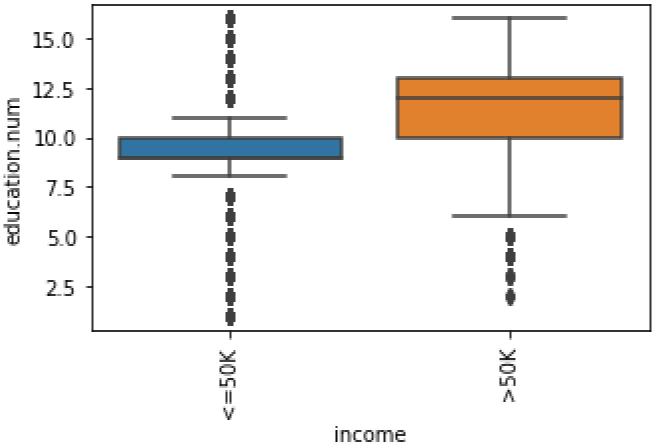


Fig-1.11: Depicts the Relationship between Education number and Income

'Education number' (education.num) refers to the number of years of education pursued by an individual. This signifies that an individual with a higher education level will be represented by a higher Education number.

As observed in the Fig-1.11, Individuals with a higher education.num have a higher income when compared to those with a low education.num thus, signifying the importance of education for higher pay.

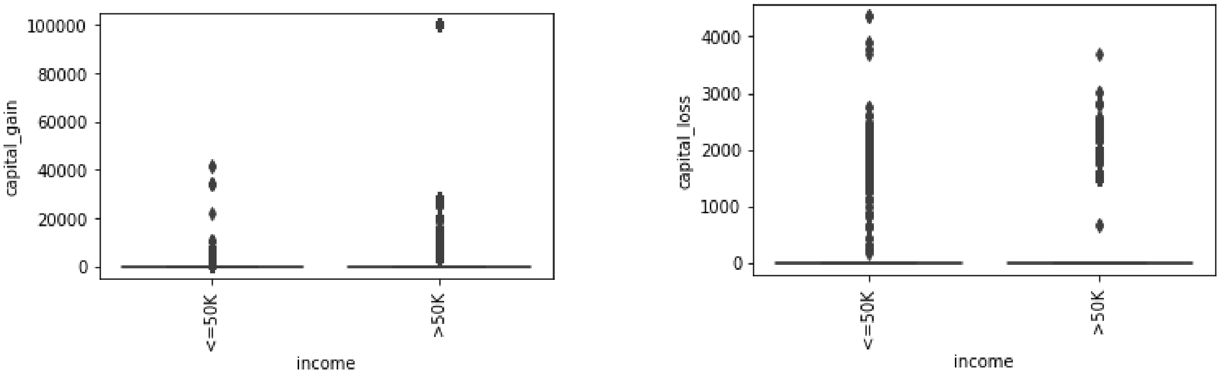


Fig-1.12: Depicts the relationship between Capital gain with income and Capital loss with Income respectively

From the above plots, no substantial relation is observed between Capital gain with income as well as Capital loss with income. We shall perform the ANOVA test for further insights.

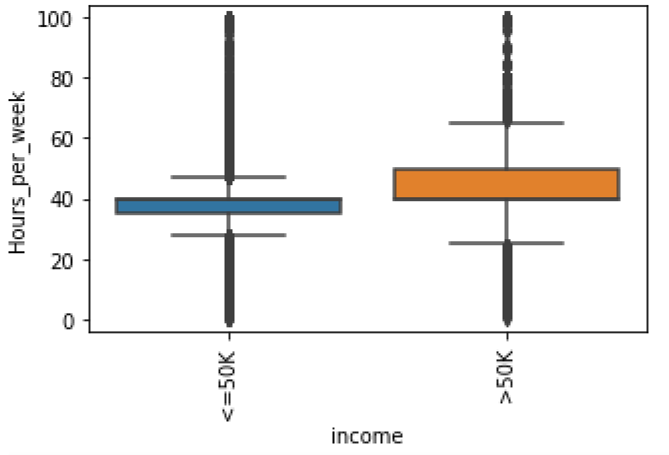


Fig-1.13: Depicts the Relationship between the Hours per week and Income

It is observed that individuals working for a longer duration tend to earn more when compared to those with less working hours.

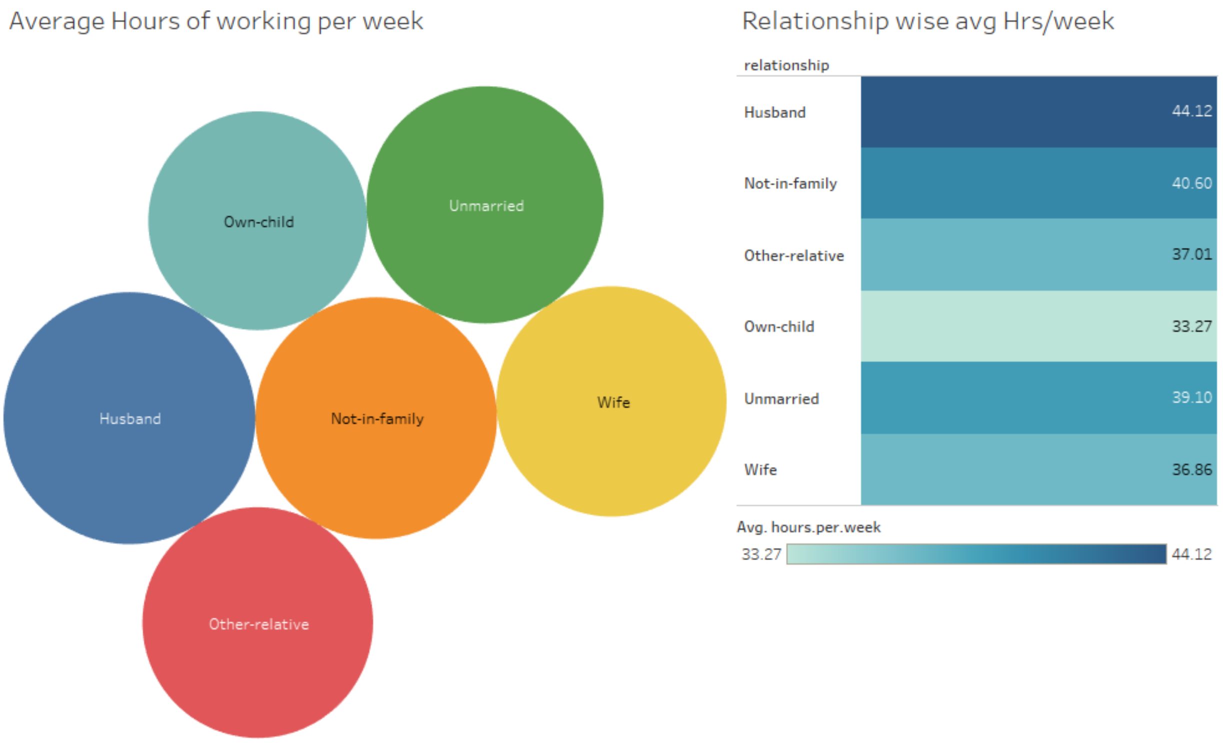
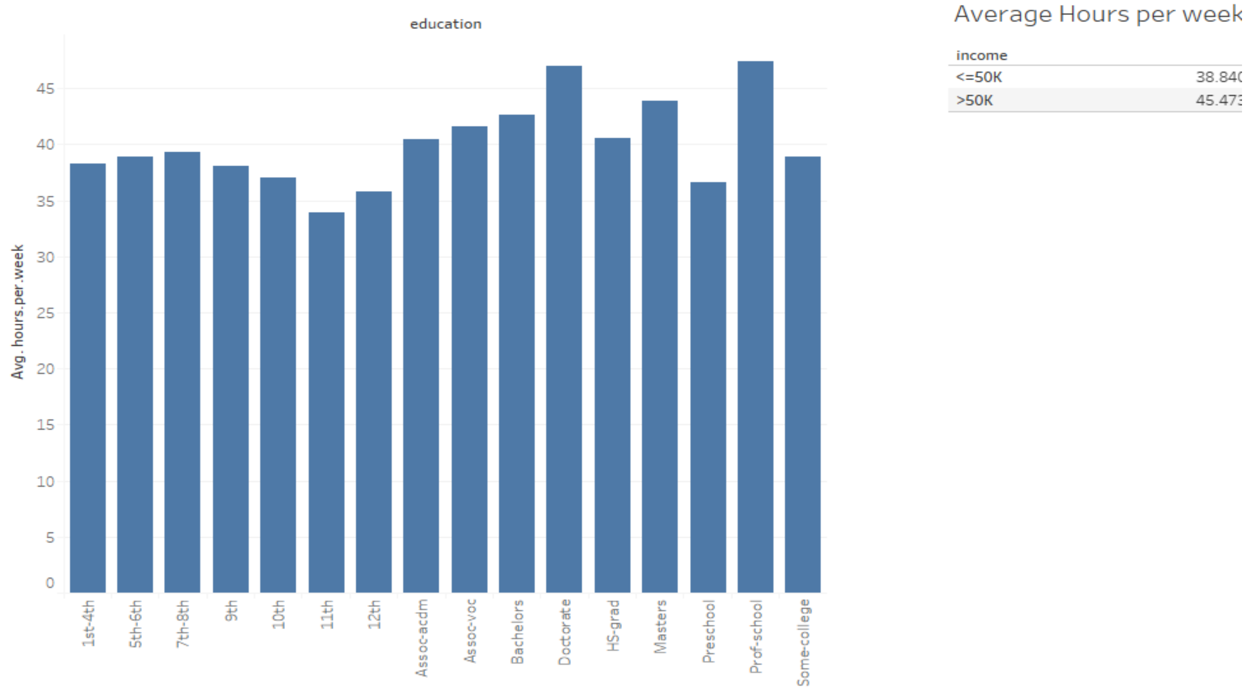


Fig-1.14: Visualizations :- Average hours of working per week , Relationship wise avg Hrs/week

The above visualizations depicts that 'Husband' work for the longest hours while 'Own-child' i.e Children below the age of 18 years work for the shortest period of hours.

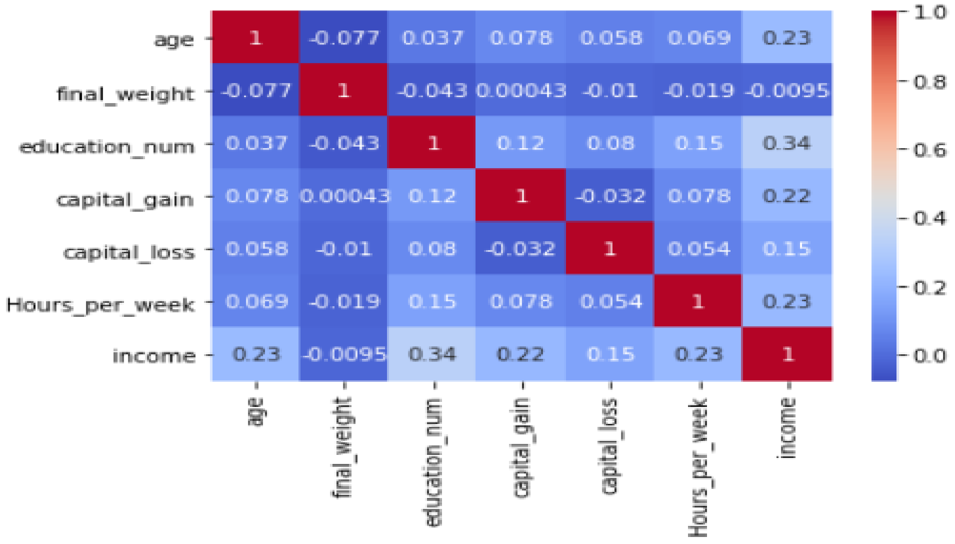


Average Hours per week

income	
<=50K	38.84
>50K	45.47

Fig-1.15: Relationship between Avg.hours.per.week and education

It is observed from Fig-1.15 that people who pursued Doctorate and School Professor courses contribute the maximum number of hours into their work which is < 45 Hours per week. This can be witnessed by their income, as they are the ones earning greater than 50K annually. Thus, it is observed that the people belonging to such classes are truly hardworking and are being subsequently paid a high salary for all their efforts.



	age	final_weight	education_num	capital_gain	capital_loss	Hours_per_week	income
age	1	-0.077	0.037	0.078	0.058	0.069	0.23
final_weight	-0.077	1	-0.043	0.00043	-0.01	-0.019	-0.0095
education_num	0.037	-0.043	1	0.12	0.08	0.15	0.34
capital_gain	0.078	0.00043	0.12	1	-0.032	0.078	0.22
capital_loss	0.058	-0.01	0.08	-0.032	1	0.054	0.15
Hours_per_week	0.069	-0.019	0.15	0.078	0.054	1	0.23
income	0.23	-0.0095	0.34	0.22	0.15	0.23	1

Fig-1.16: Correlation of variables with Target using Heat map

Inferential Statistic

Chi-square test for categorical vs categorical variables

For every statistical test there goes a null and Alternate Hypothesis side by side and then we move ahead towards testing of the variables. In chi2 contingency test the null Hypothesis is that the two variables are dependent and have no significant difference between variables where as Alternate as name suggests is opposite of null and it suggests that the variables are independent and hold no relationship with each other. The test returns values of chi2, p, dof and expected in array form.

If p-value is less than alpha this means that we reject the Null Hypothesis means that there is a significate difference between the two variables.

After checking for all categorical variables one by one in our dataset all variables passed the test and hence we did not drop any variable based on this test.

Anova Test for Categorical vs Numerical variables

For Anova null hypothesis goes as follows-

H0: The mean is same for all groups

H1: The mean of atleast one of the variables differ

If p-value is less than alpha this means that we reject the Null Hypothesis means that there is a significate difference between the two variables. We saw that the p-value is less than alpha(0.05) for all variables except **'fnlwgt'**, So we dropped 'fnlwgt' from our dataset.

Correlation

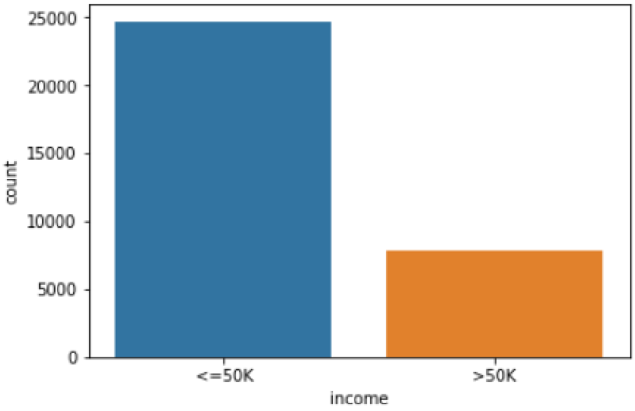
- 'Education' and 'Education Number' has a very high correlation thus one of the two columns can be drop to remove muticollinearity
- There is no relationship between these two variables thus we don't drop any of the column.
- There is relationship between 'Marital Status' and 'Relationship'. Thus one of the two columns can be drop to remove muticollinearity

Techniques to Balance dataset

There are several techniques to balance the dataset if highly imbalance classes found in target. Undersampling, Upsampling and Smote all work differently and can be used according to our requirement.

Smote

Synthetic Minority Oversampling Technique is a statistical technique we used to balance our data by oversampling our minority class. It uses K nearest neighbors' technique to synthetically modify the training data.



We used SMOTE because Income which is our target had imbalanced classes in our dataset with "<=50k" having 76% data and ">50k" only 24%. Having a balanced set will make our dataset less prone to biased classification and better accuracy.

Learning Algorithm

We used many algorithms and one of the Ensemble techniques, Random Forest Classifier gave us the best results, so considering it as our Learning algorithm to predict the Income class. It's a Bagging technique which tries to make a weak learner from all the strong learners/overfit models, typically decision trees.

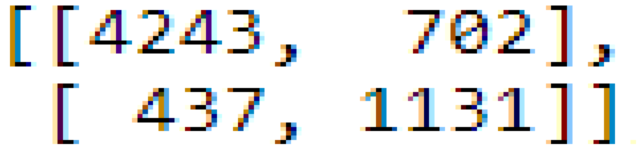
We get 83% accuracy and F1 Score achieved is:

$$\mathbf{F1} = \frac{\mathbf{2}}{\left(\frac{\mathbf{1}}{\mathbf{recall}}\right)+\left(\frac{\mathbf{1}}{\mathbf{precision}}\right)} = \mathbf{83.2}$$

Structure of Confusion Matrix: [Table 1]

	Predicted		
Actual		class<=50k	class>50k
	class<=50k	True Negatives	False Positives
	class>50k	False Negatives	True Positives

Confusion Matrix of Random Forest:



```
[[4243,  702],
 [ 437, 1131]]
```

By observing we found out that False Positives are comparatively higher than False Negatives that means people who are under 50k have chances of getting misclassified in the above 50k range.

Comparison with Base Model: [Table 2]

Model	**Accuracy Score**	**Precision Score**	**Recall Score**	**F1 Score**
Linear regression(Base)	79.01	77.4	79.0	75.5
Random Forest (Final)	83	85.5	82.4	83.26

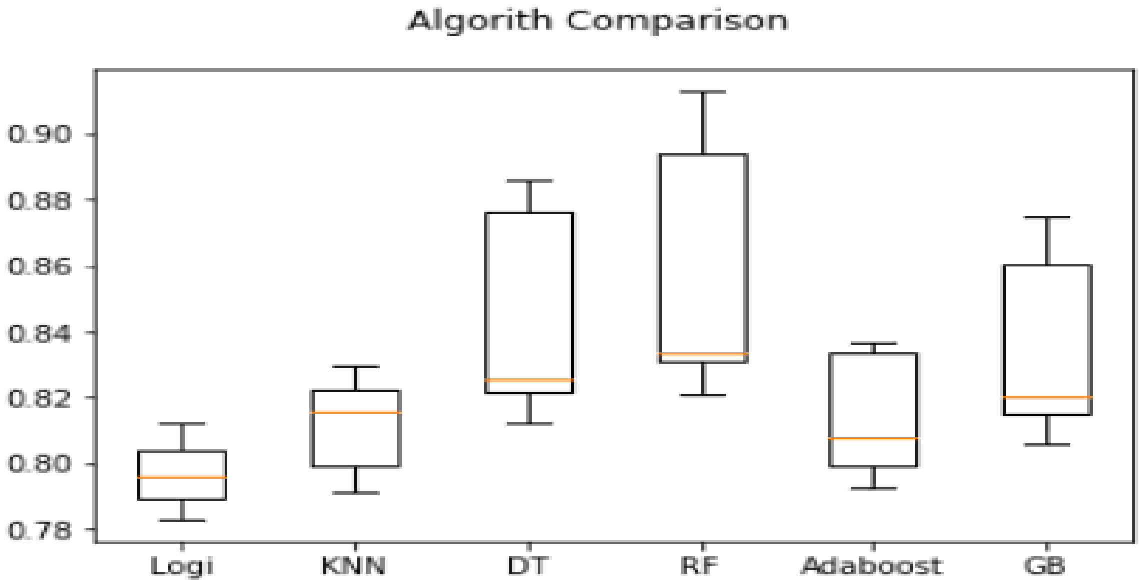


Fig-1.17: Box Plots showing the best fit models for our dataset with accuracy on y-axis and name of algorithms on x-axis.

AUC-ROC Curve

Area under the receiver operating characteristic curve, shows the ability of classifier as its threshold varies. It is a plot of True Positive Rate/Sensitivity/Recall against False Positive Rate. False positive rate can also be denoted as 1-Specificity.

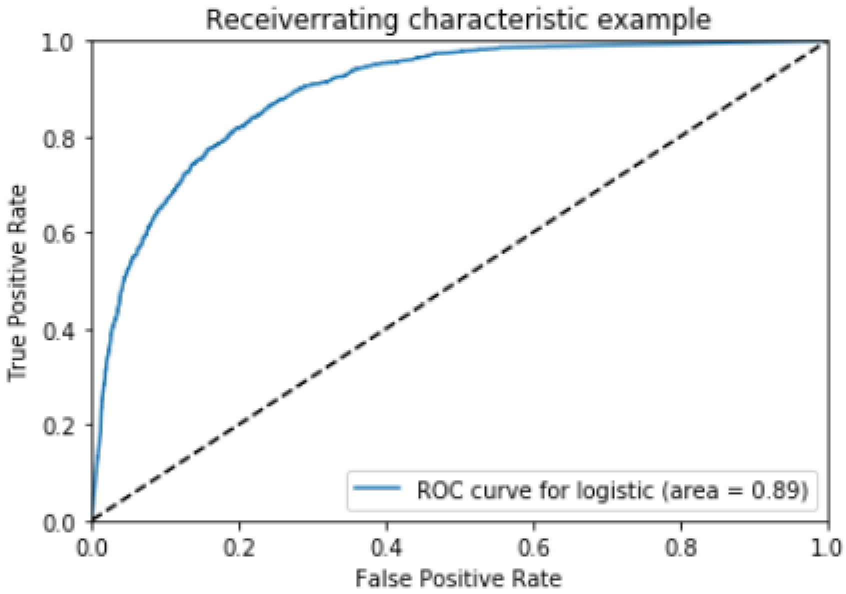


Fig-1.18: Shows 89% of area under the ROC curve

Conclusion

1. We can see that the column name ‘fnlwgt’ has no impact on the income level of the individual thus we can drop this column.
2. The ‘race’ column is also biased towards white so this column is also not giving additional information to the model. Thus, we can drop this column
3. The ‘Native Country’ is also biased towards the US so this column is also not giving additional information to the model. This, we can drop this column.
4. We are dropping ‘Relationship’ and ‘Education’ as they are highly correlated with other columns.
5. After all this we converted less than $50K income 0 and more than $50K income 1 for the machine building model.
6. Random Forest Classifier, a Bagging technique works best for this dataset.

References

[1] UCI machine Learning: "*Adult Census Income - predict weather income exceeds$50k/year based on census data*", https://www.kaggle.com/uciml/adult-census-income

[2] Navoneel Chakrabarty and Sanket Biswas: "*A Statistical Approach to Adult Census Income Level Prediction*", https://arxiv.org/ftp/arxiv/papers/1810/1810.10076.pdf

[3] Vidya Chockalingam, Sejal Shah, Ronit Shaw: *"Income Classification using Adult Census Data"*,https://pdfs.semanticscholar.org/3dd5/e9f335511efbb81d65f1d6d4995019f8b5fd.pdf

[4] *"Building your First Neural Network on a Structured Dataset"*,https://medium.com/analytics-vidhya/build-your-first-neural-network-model-on-a-structured-dataset-using-keras-d9e7de5c6724

[5] *"Application of Synthetic Minority Over-sampling Technique (SMOTe) for Imbalanced Datasets"*,https://medium.com/towards-artificial-intelligence/application-of-synthetic-minority-over-sampling-technique-smote-for-imbalanced-data-sets-509ab55cfdaf