# Income Prediction

Presenter: Aaqib Ali

# AGENDA

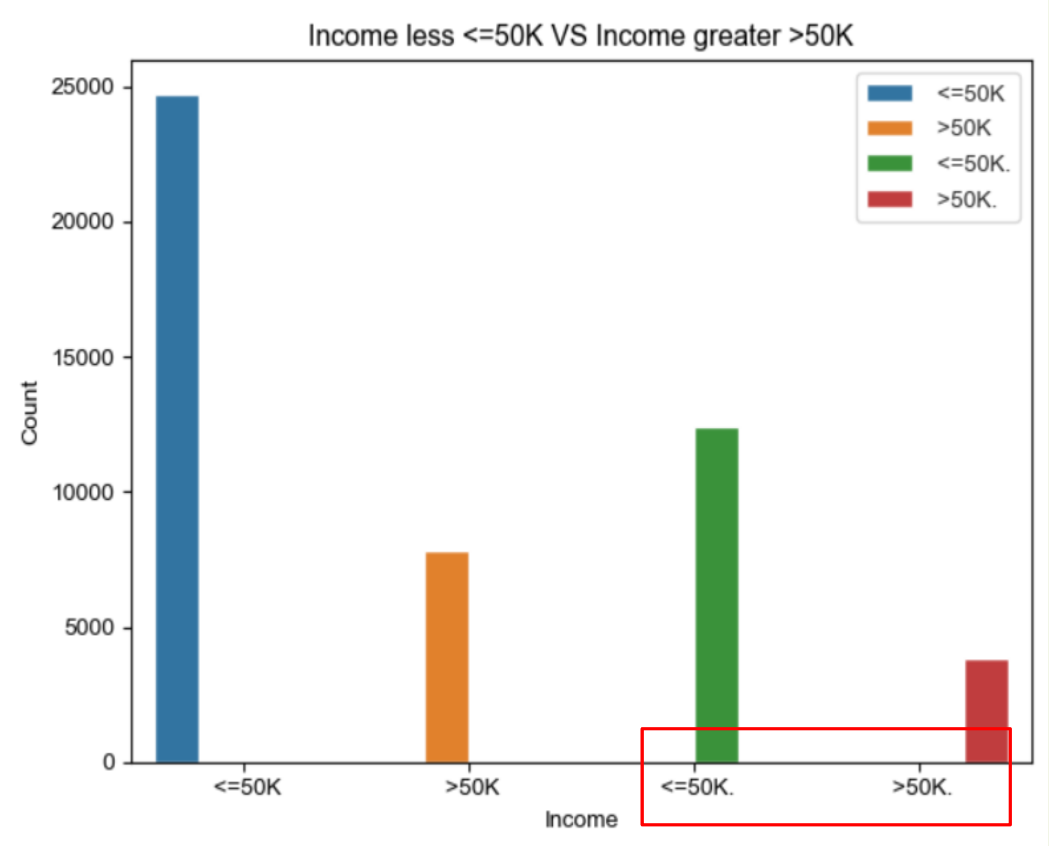
- Introduction
- Data Analysis and Exploration
- Model Approach
- Model Evaluation Metrics
- Model Performance
- Driving Factors Analysis
- Model Deployment
- Conclusion

# INTRODUCTION

- In this task I have to analyse the given “Census” dataset and build a predictive model that can predict wether a person can earn more than 50k or not. Moreover, highlight some of the driving factors of the prediction.
- Also, point out some of the techniques for model deployment.

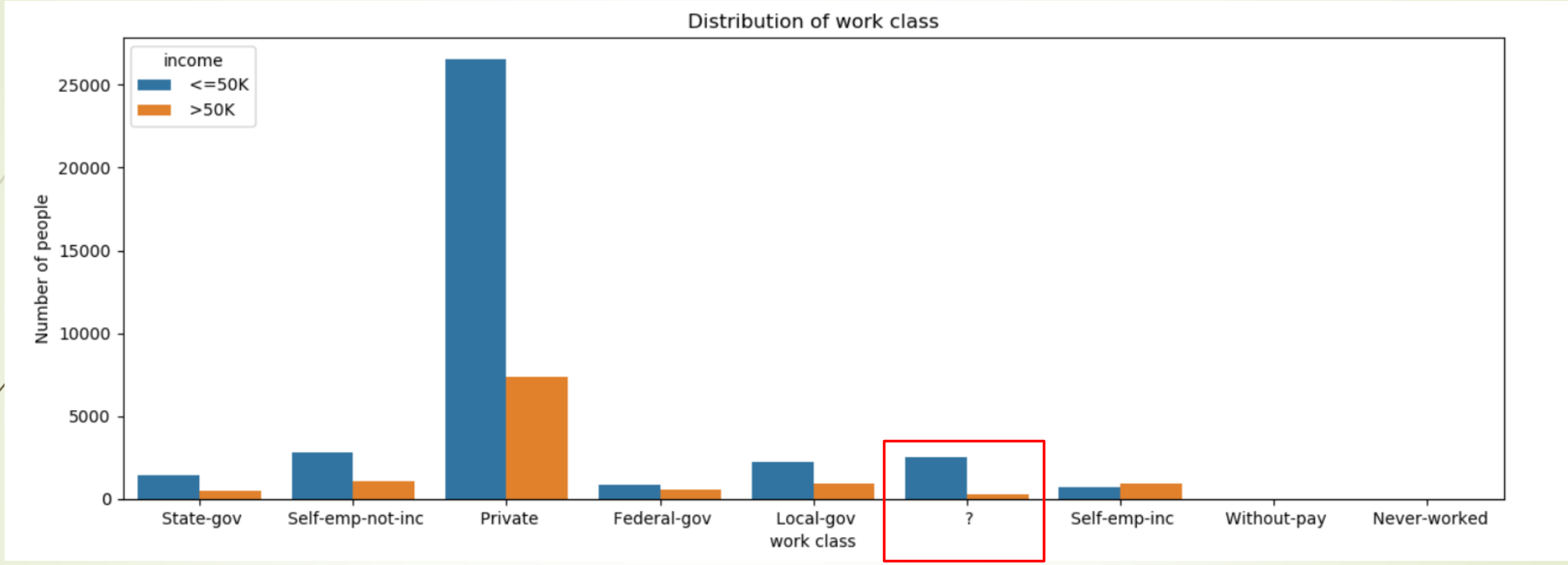
# DATA ANALYSIS AND EXPLORATION

- Data exploration and analysis is mostly used in order to understand the distrubution of the data.
- Moreover, to detect the outliers in the dataset.
- We came to know that our target variable is inconsistent.



5

# DATA ANALYSIS AND EXPLORATION

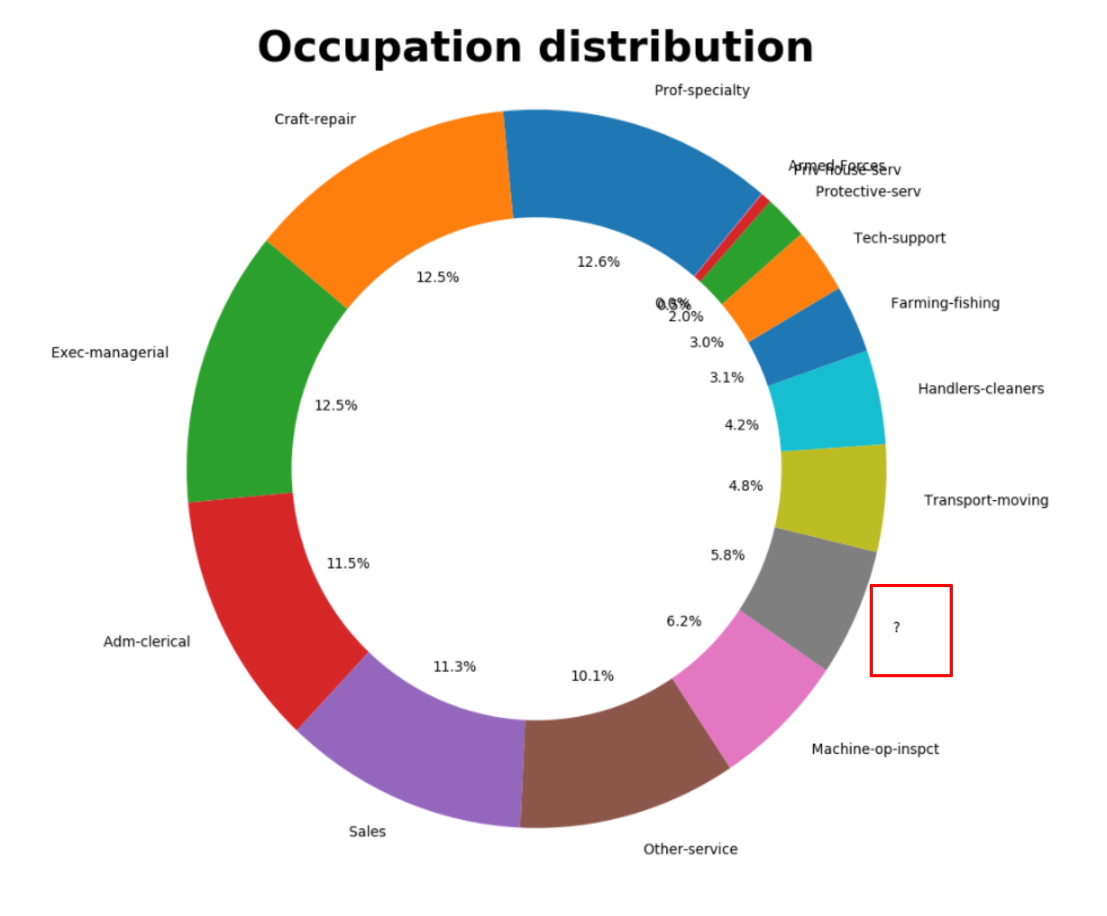


- Most of the population in our dataset is for private individuals.
- We can also spot an outlier in this columns that need to be pre-processed.

# DATA ANALYSIS AND EXPLORATION

- Around 60% of the occupation is made of Prof-Specially, Craft-repair and sales etc.

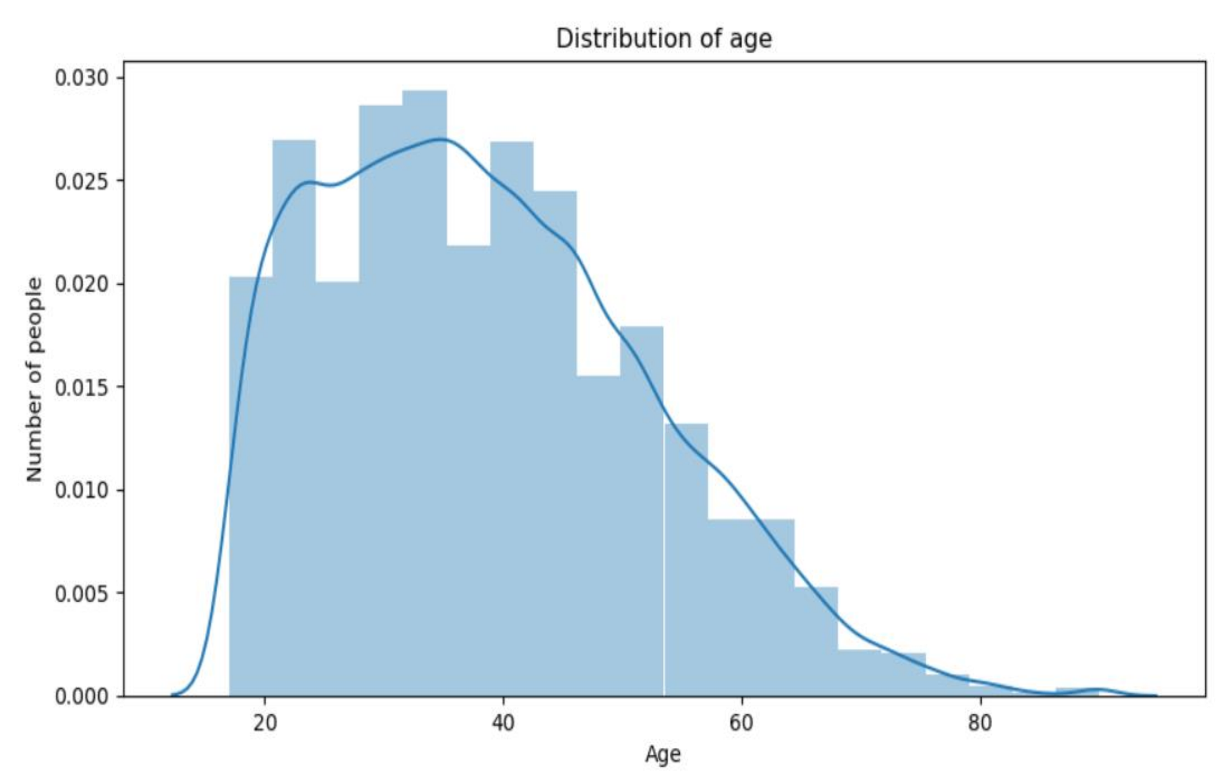
- We can also spot an outlier in this columns that need to be pre-processed.



7

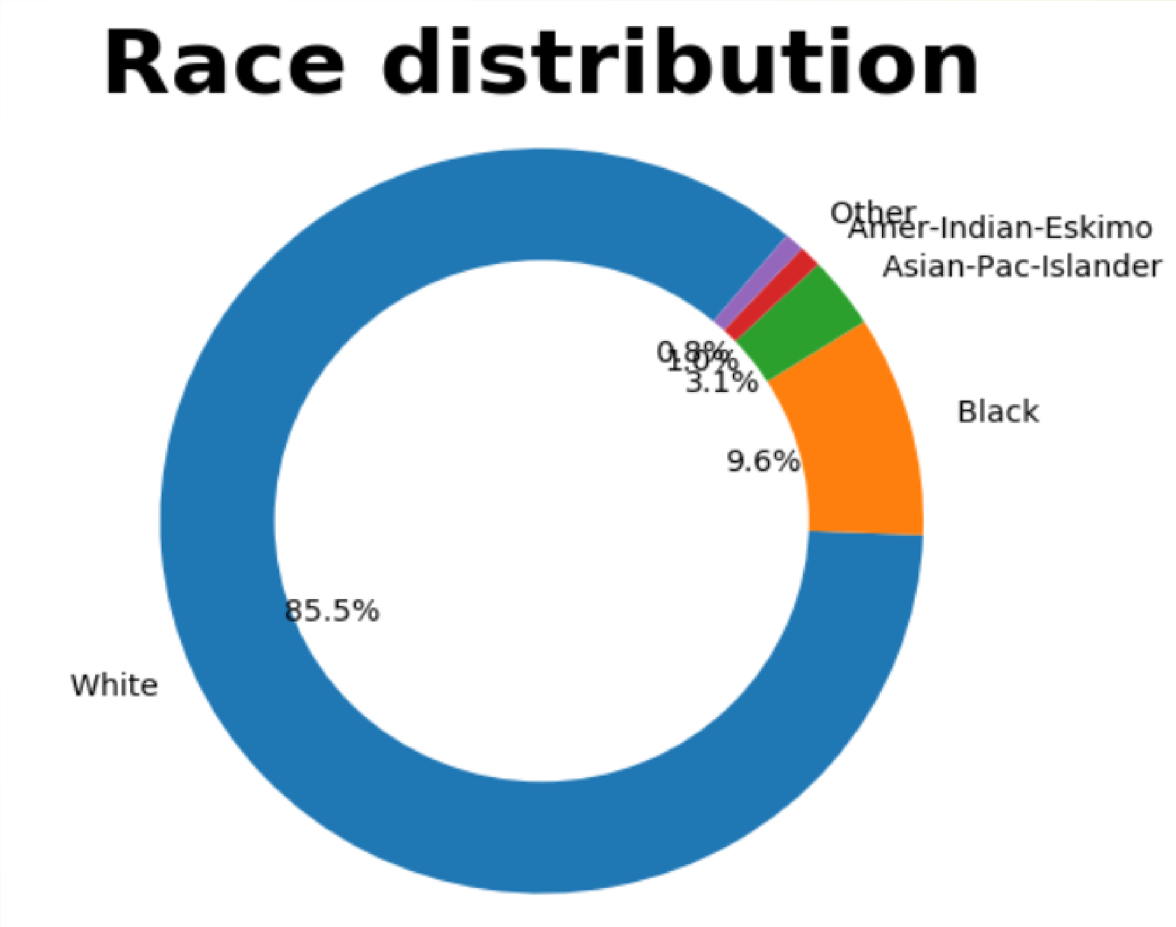
# DATA ANALYSIS AND EXPLORATION

- Most of the distribution of age lies in our dataset is between 25 to 45 years



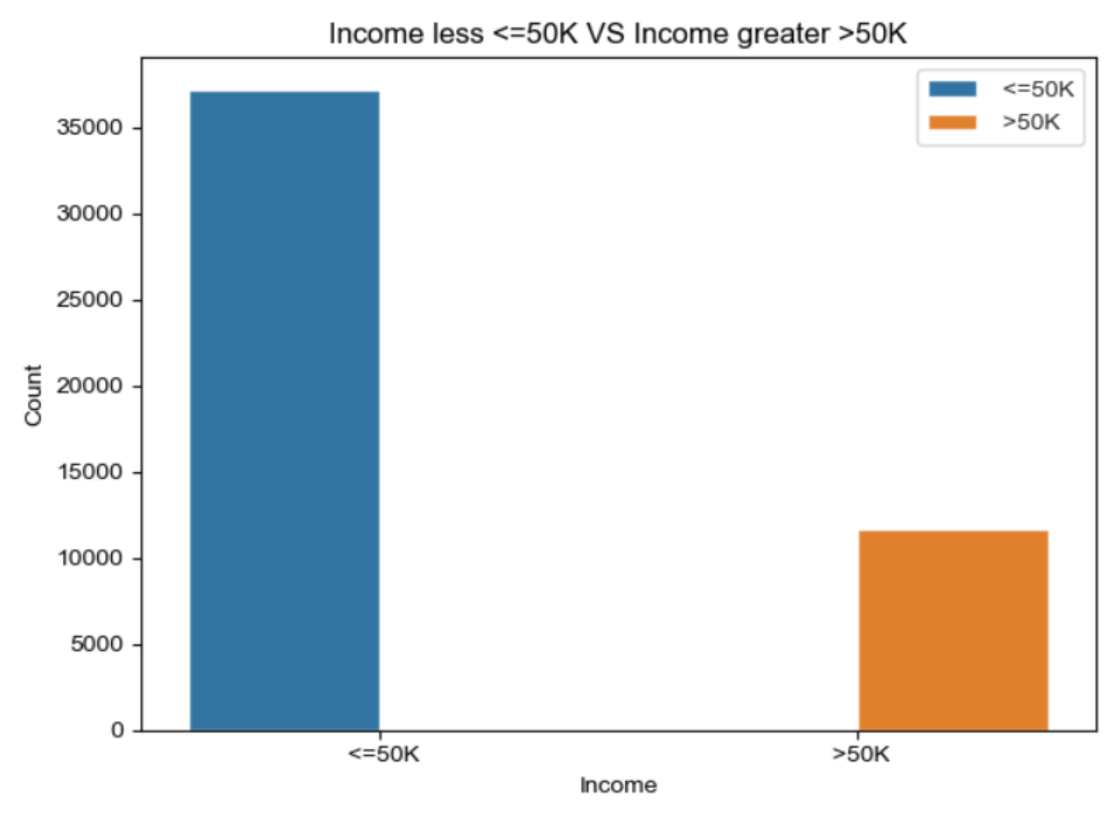
# DATA ANALYSIS AND EXPLORATION

- Around 35% of the race distibution is made of White's



# DATA ANALYSIS AND EXPLORATION

- We have less number of Income greater than 50k that is obvious.
- As, number of Income greater than 50k is less, that leads us to a common Machine Learning problem of  data imbalance.
- E.g. fraud detection



# MODEL APPROACH

- LogisticRegression is used as a baseline model. The performance is improved by using tree based models i.e. Random Forest and XGBoost.
- Random Forest and XGBoost are highly interpretable and easy to understand as compared to regression based models.
- If you want to explain the model performance and working to the business people, then these tree based models are the best option to work with. This is the reason I am working with these models.
- I achieved best performance and AUC with XGBoost while LogisticRegression performs quite low.
- To deal with the data imbalance problem I used a weighted loss technique which is considered to be the most effective Machine Learning technique as compared to down sampling and over sampling.
- For features selection, I extensively used feature importance, panda profile report and L1 regularization (Lasso Technique).

# MODEL EVALUATION METRICS

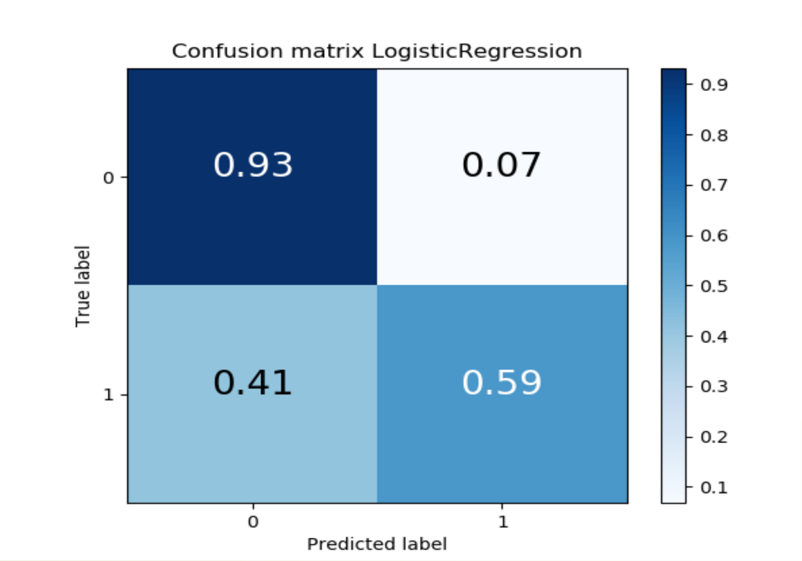
- As, we are dealing with the supervised classification problem, for this type of problems “Accuracy” is the most suitable evaluation metrics.
- However, we have a data imbalance problem, in our case “Accuracy” metrics will not work.
- We will get 90% accuracy as our model is predicting negative class every time and unable to classify the Positive class at all. Then this 90% accuracy is actually wrong.
- For our special case, I used “Area under the curve” metrics which is the suitable metrics for imbalance datasets.
- Our main goal is to Improve the “Precision“ of the negative class (Income <=50k) as well as “Recall” of the positive class (Income >50K).
- For analysing the driving factors I am using “SHAP Package“ from python.

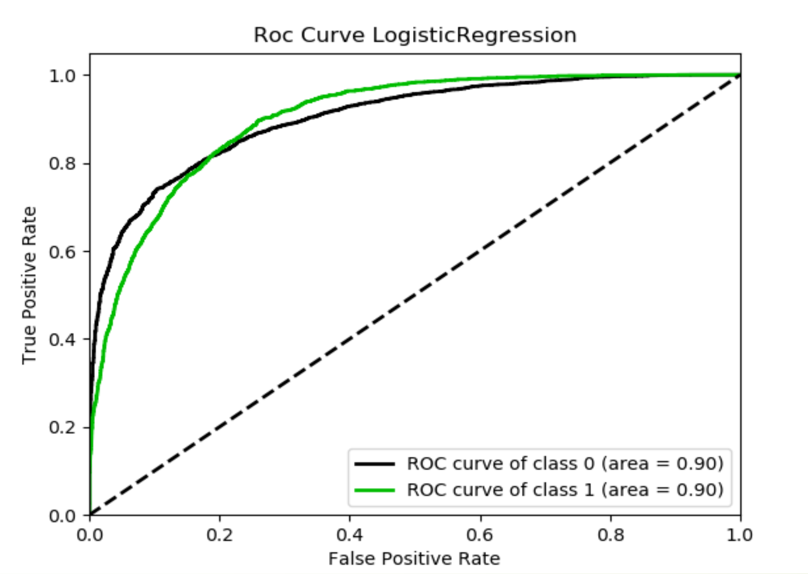
# MODEL PERFORMANCE AND RESULTS

- XGBoost is trained on 80% of the final dataset and 20% of the dataset was used to validate/test  the performance of the model.
- Trained LogisticRegression, Random Forest and XGBoost. However, final submission contains the best performed model i.e. XGBoost.
- All the model hyper parameters are also tuned and final model is trained on the best parameters.
- Final result contains ROC curve, Confusion Matrix, SHAP analysis and model deployment.

# MODEL PERFORMANCE

## LogisticRegression with weighted loss technique:

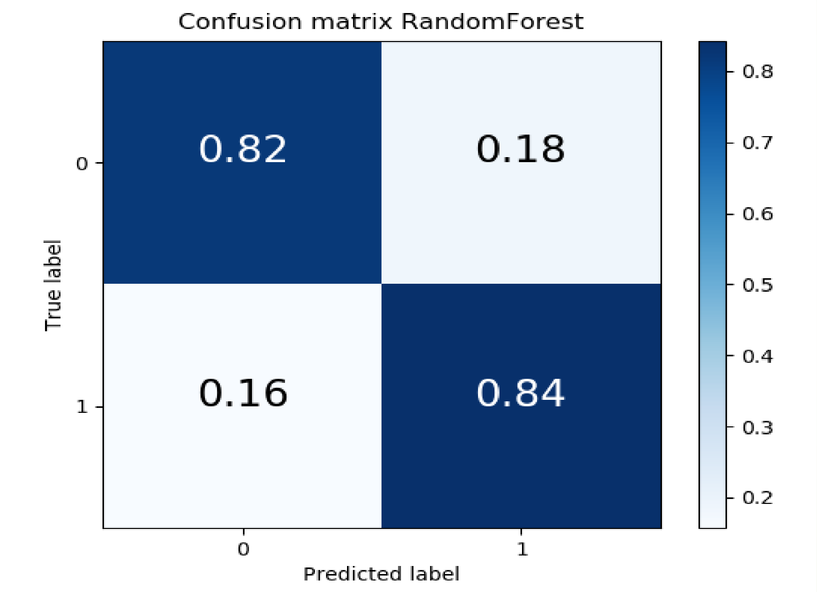


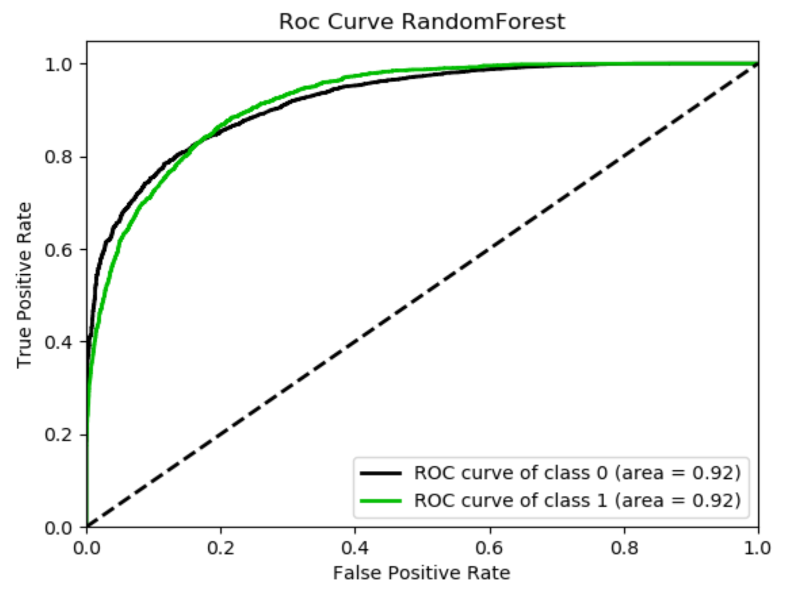


- LogisticRegression is covering 90% of the area under the curve (AUC) and also predicting both classes  perfectly. However, some of the negative labels are wrongly classified which needs to be  improve.

# MODEL PERFORMANCE

## Random Forest with weighted loss technique:

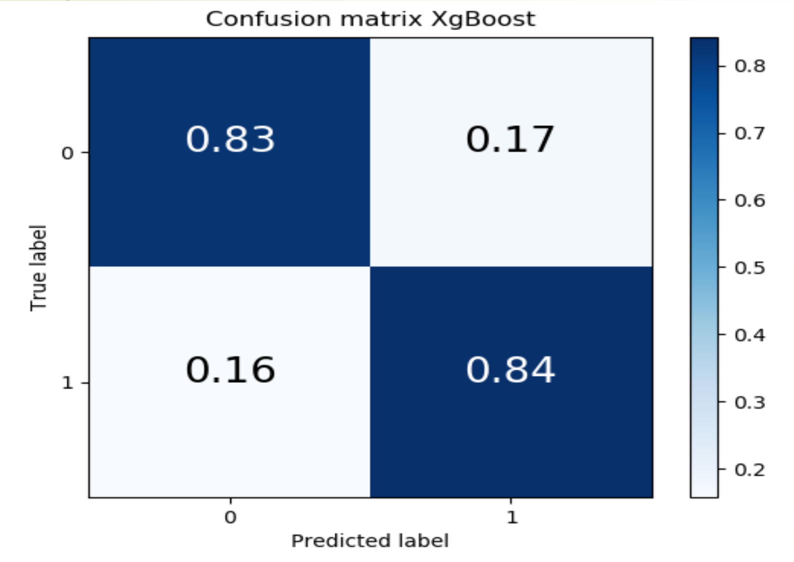


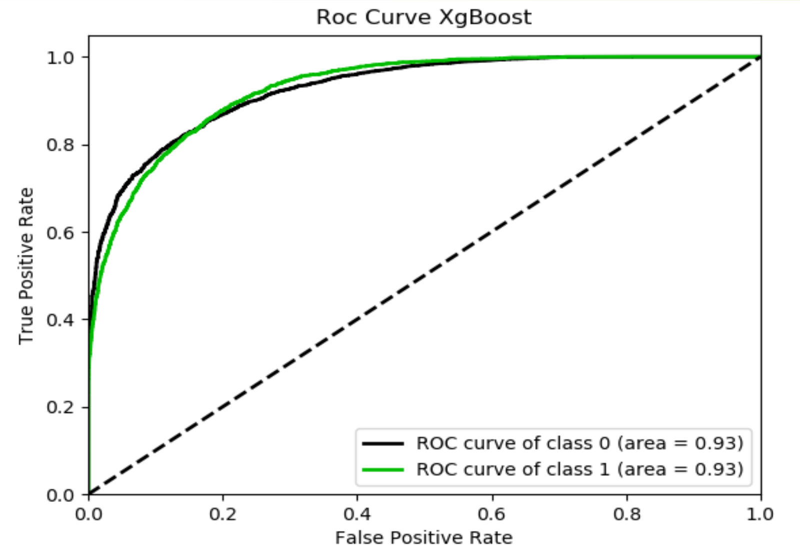


- Random forest is covering 92% of the area under the curve (AUC) and also predicting both classes  perfectly. However, we can try further to improve the AUC.

# MODEL PERFORMANCE

## XGBoost with weighted loss technique :

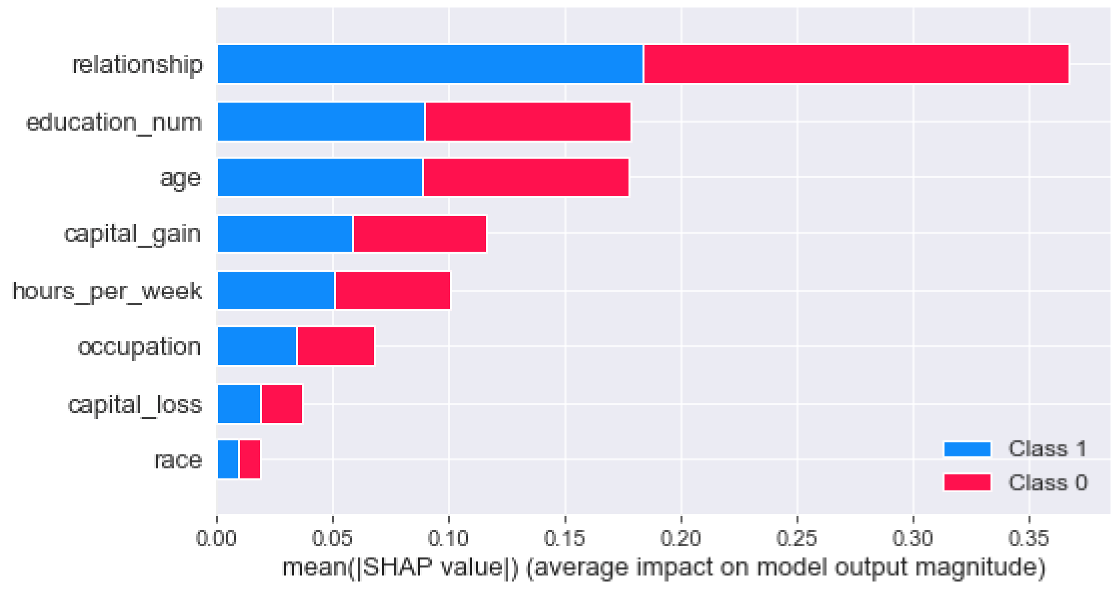




- XGBoost is performing the best so far on the dataset by covering 93% of the area under the curve. Moreover, model is predicting both classes perfectly.

# DRIVING FACTORS ANALYSIS

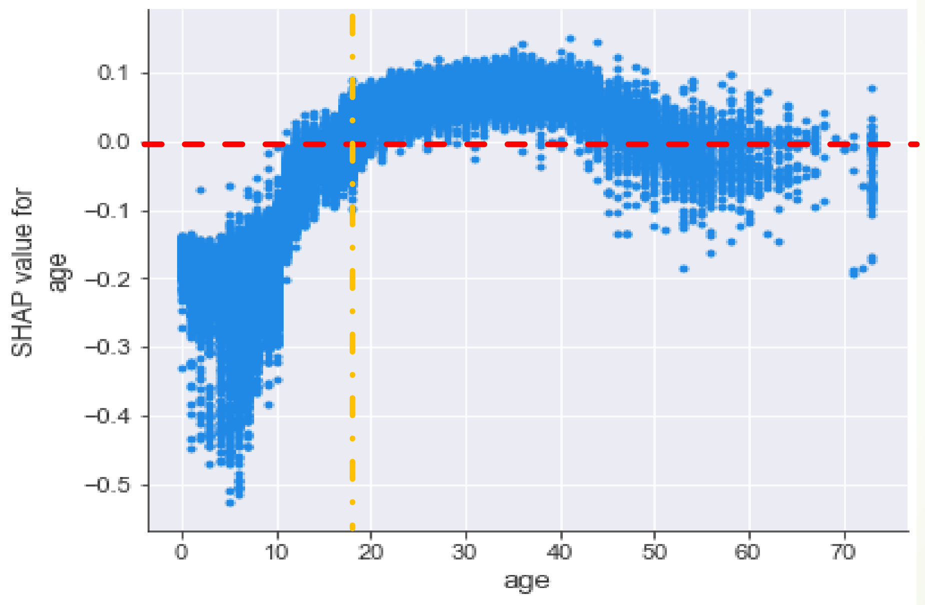
## Shap features importance:



- Relationship, number of education years and age are the top three influencing factors for the model prediction. We can further analyse these factors by their shap explaination.

# DRIVING FACTORS ANALYSIS

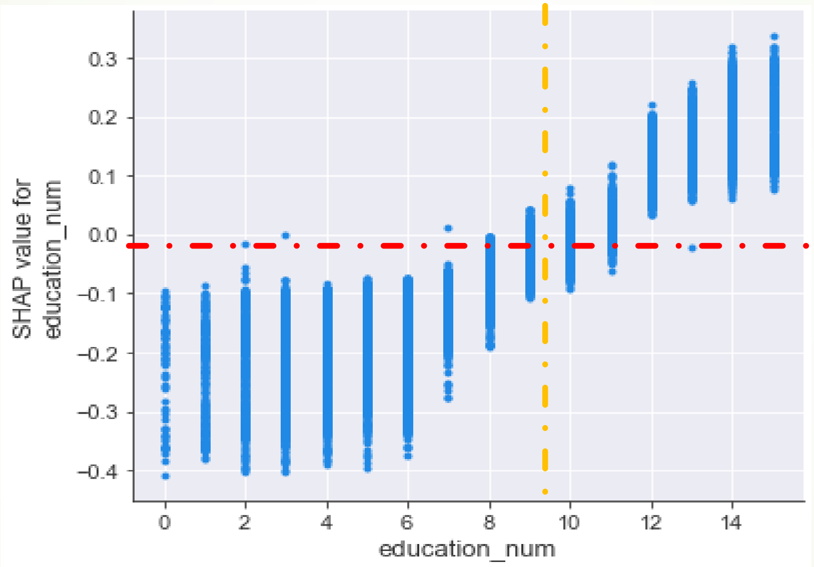
**Age dependency plot:**



- From 18 to 40 years of age, It has a positive effect on the prediction.

# DRIVING FACTORS ANALYSIS

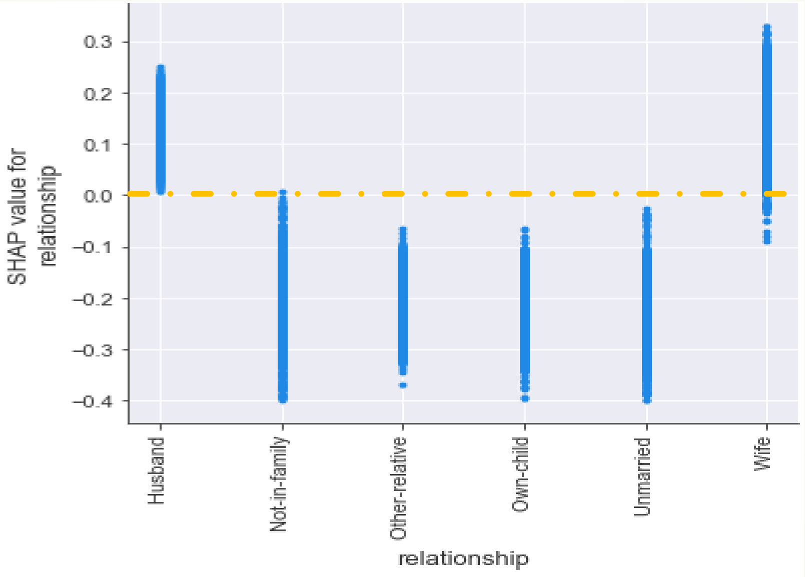
## Education years dependency plot:



- We can clearly see a significant lift after 10 years of education that has a postive influence on the prediction.

# DRIVING FACTORS ANALYSIS

## Relationship dependency plot:



- We can clearly from the graph that relation of husband and wife have a positive impact on the prediction as compare to the other relationships.

# MODEL DEPLOYMENT

- There are many ways to serve the model in production.
- Model deployment always depands upon stakeholders.
- Some of the techniques listed down:
    - Flask deployment with web interface. (Included in the Implementation)
    - API endpoint.
    - Cloud deployment. E,g. EC2 instance.
    - Cloud Infrastructure E.g, Kibana deployment (Model monitoring)

# CONCLUSION

- Mode performance is in green area as we are acheiving 93% of AUC,
- Model performance could be improved but it will be very hard to improve that and also rate to improve will be also low.
- We see from the SHAP analysis that Relationship, age and number of education years are the top influenceing factors for the model.
- We further analyze these driving factors by SHAP explainer.
- I choose XGBoost my preferred approach as we are getting high AUC with it.
- XGBoost is not only simple but also easy to interpert by the stakeholders with the help of features importance and SHAP values.
- It also reduce the chance of overfitting.

Thank You.... !!!!