# Prediction of the salary class of a person from the census data

By Nanda K

# Overview

- Introduction
- Data
- Exploratory data analysis
- Modeling the data
- Conclusion
- Next steps

# Introduction

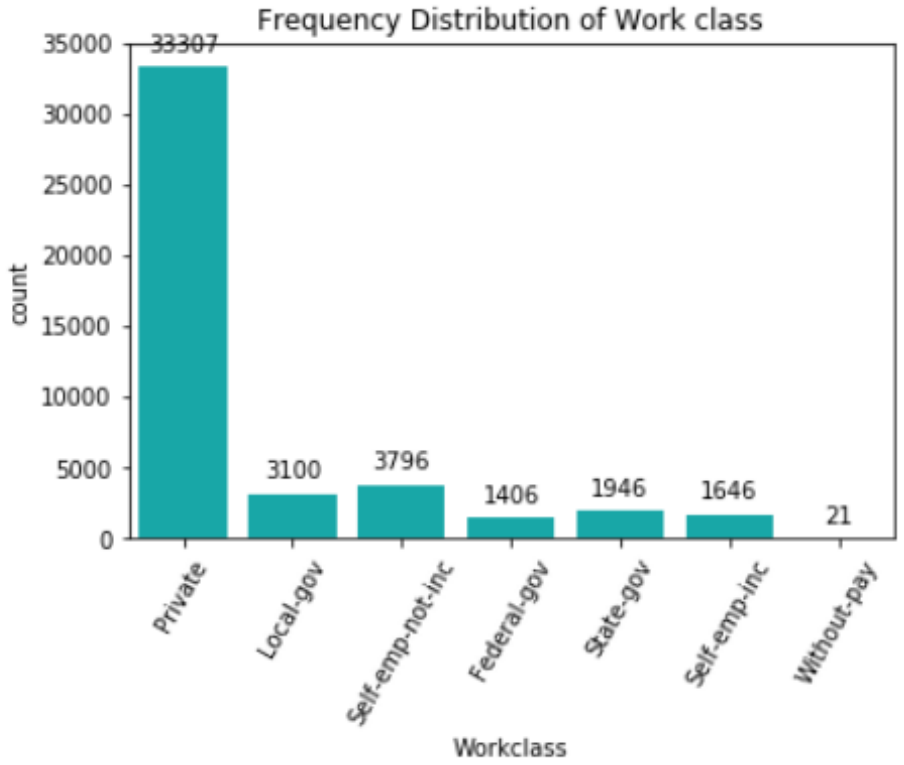
Filtering the population based on salary class is helpful to marketing campaigns for signing up to educational programs, selling of automobiles, to find clients for banking products like term deposits, credit cards etc.

With the given set of attributes from the census data, which are Age, Work class, Final weight, Education, Education-number, Marital-status, Occupation, Relationship, Race, Sex, Capital-gain, Capital-loss, Hours-per-week, Native-country, we have to predict if the person's salary class belong to >50k per year class or <=50k per year class.

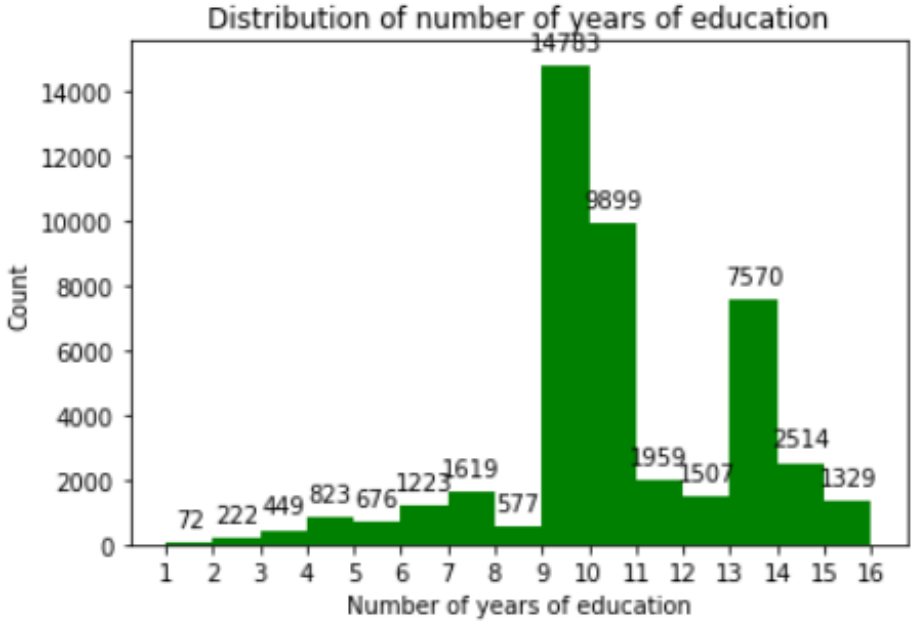
# Data

- Dataset consists of 15 attributes which are Age, Workclass, Finalweight, Education, Education-num, Marital-status, Occupation, Relationship, Race, Sex, Capital-gain, Capital-loss, Hours-per-week, Native-country, Class. It contains continuous and nominal attributes. It has 48842 instances of data with some missing values.
- The dataset was extracted by Barry Becker in 1994 from census data of the United States.
- After Data wrangling, dataset has 11 features and target variable, Class.

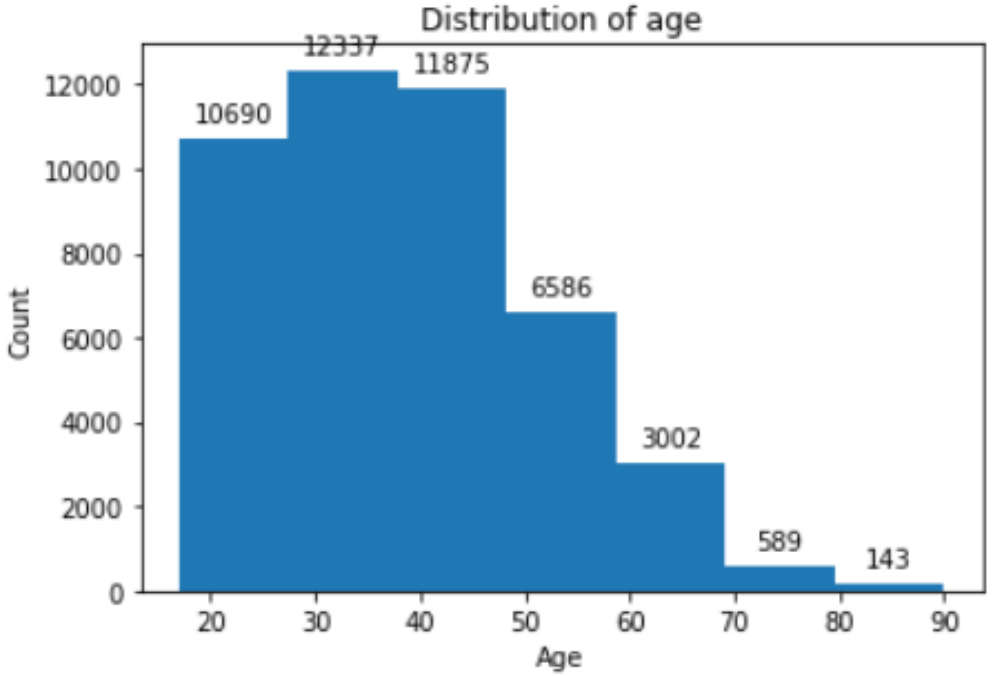
# Exploratory Data Analysis




Majority of the given population works for Private sector. There are 21 people without pay.

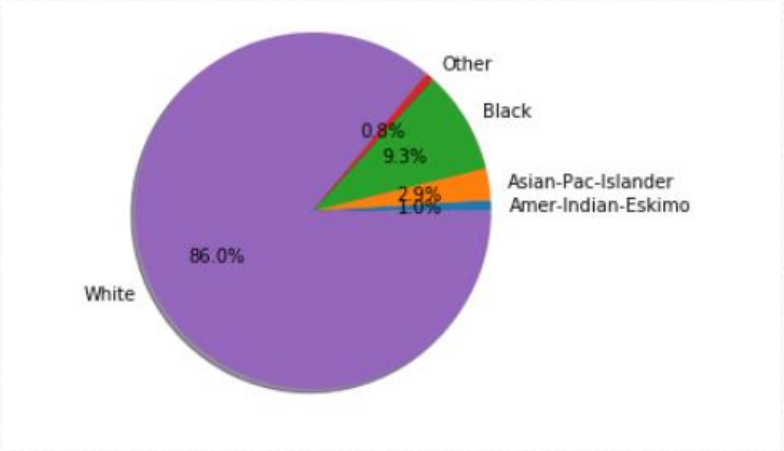



The number of years of education in the dataset ranges between 1 - 16 years. Maximum number of people have 9 years of education.

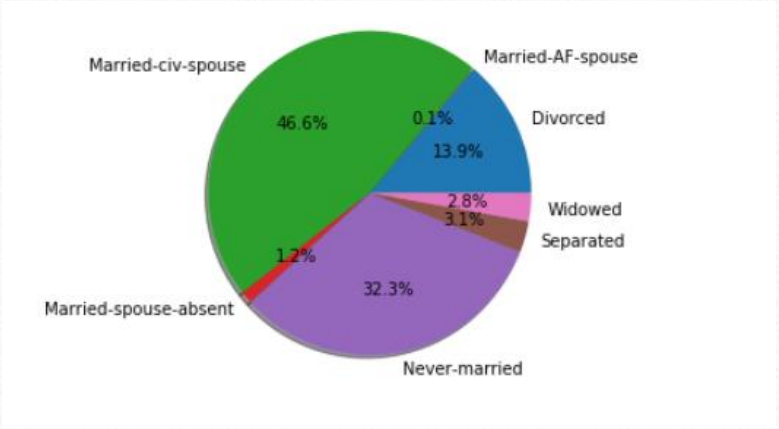



From the plot above, we can see that maximum number of people in the dataset are between 30-40 years and least number of people in the group 80-90 years.

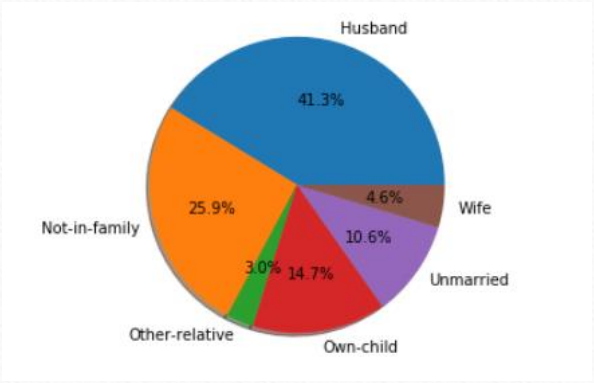

Average number of hours worked in the dataset is 40.9 hrs. There are a number of outliers which are number of hours worked less than 32hrs and more than 51hrs which do have predictive power in predicting the income class of the person.



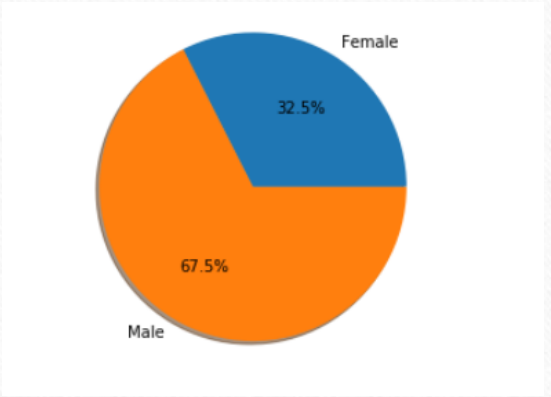

Distribution of race




Distribution of marital status

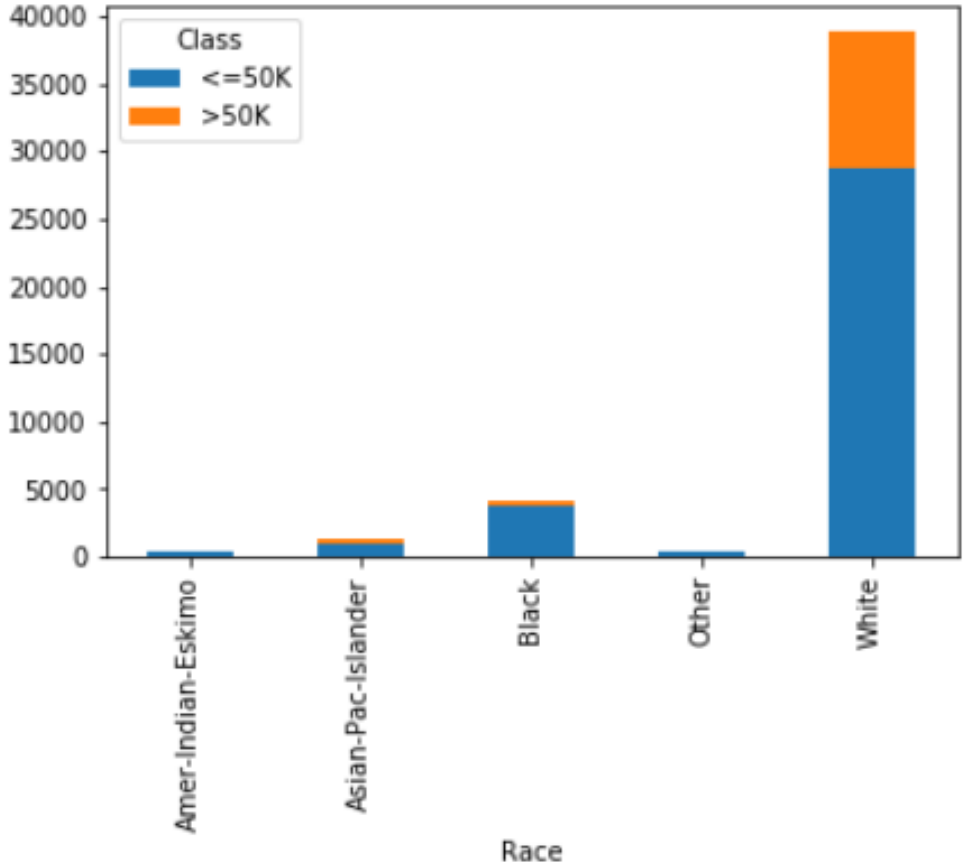



Distribution of relationship



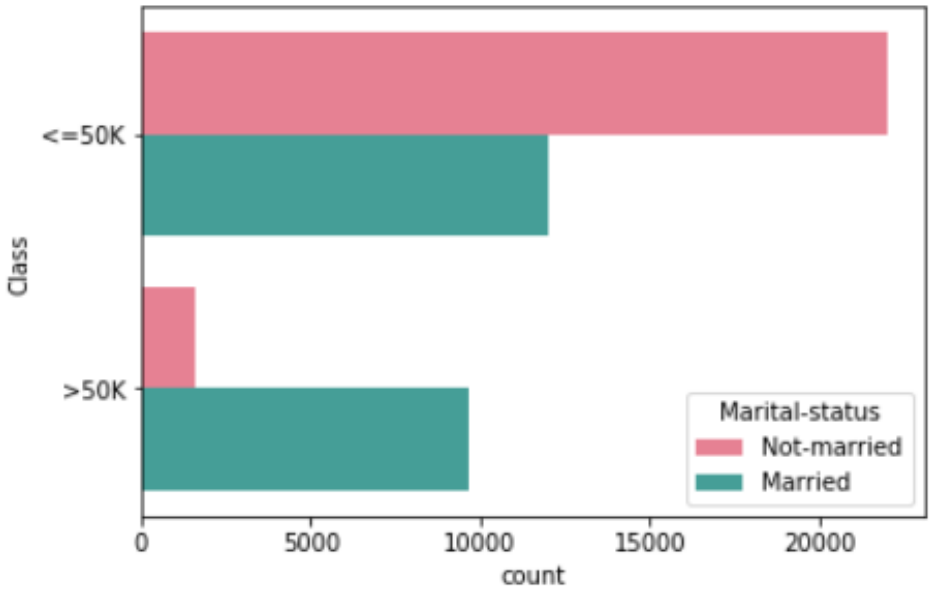

Distribution of sex

**Who are the majority in the >50k income class?**



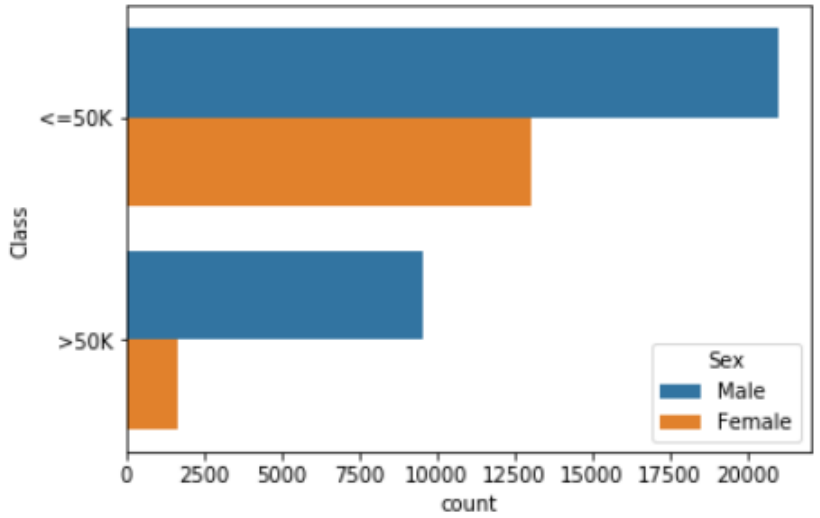

The white race seem to have dominated the income class of more than 50k. But then the given population itself is dominated by white.

**Does the income class of >50k have more married people or not married people?**



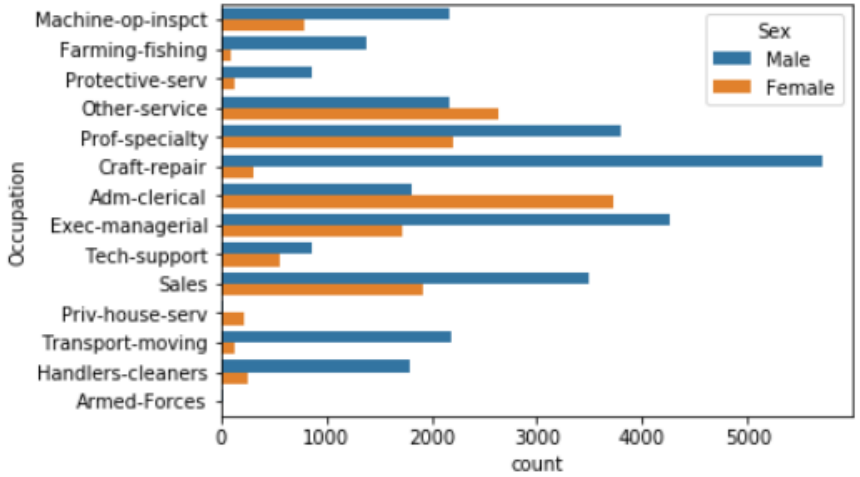

Majority of the income class of >50k has married people. But the income class of <=50k has more Not-married people than married people.
Maybe the responsibility of being married and having to maintain the household will push the person to earn more.

**Does the income group of >50k have more men or women ?**




Men comprise of 2/3rd and women of 1/3rd the given population.
1/3rd of the population is women, but only 1/8 of the income class of >50k is women. Most of the people in the income class of >50k are men.

**Distribution of men v/s women in various occupation**




There are significantly more men than women in transport-moving occupation, which can be explained by the need of physical strength for that
occupation. Admin-clerical occupation has more women than men.

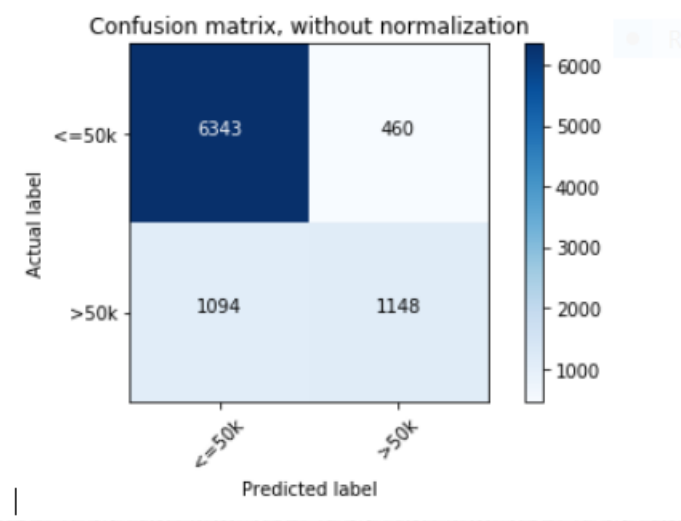
# Modeling the data

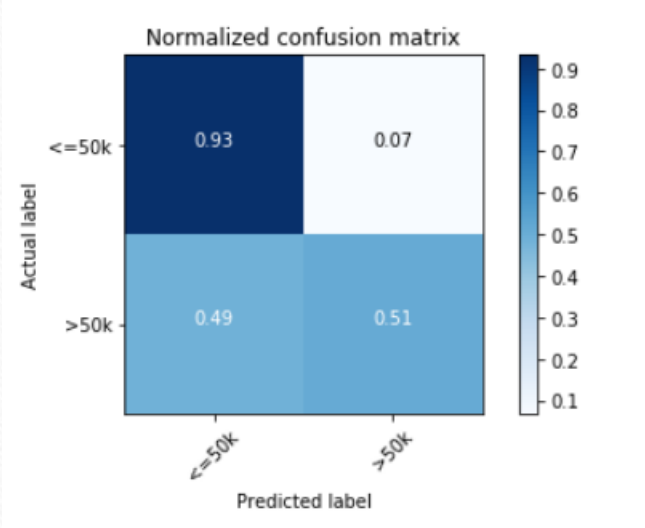
- Modelling the data on Support Vector Machine and Random Forest Classifier algorithm.
- Support Vector Machine considered because of relatively clean and less noisy data.
- Random Forest Classification Algorithm considered because dependent variables are mix of categorical and numeric values.
- Evaluating the performance of the model by Confusion matrix and Area under the ROC curve.
- Confusion matrix assess the accuracy of the predicted model, gives the number of True positive, False Positive, False Negative, True Negative cases.
- AUC denotes how well the model is distinguishing between classes. Higher the AUC, better the model is distinguishing between classes

## Support Vector Machine

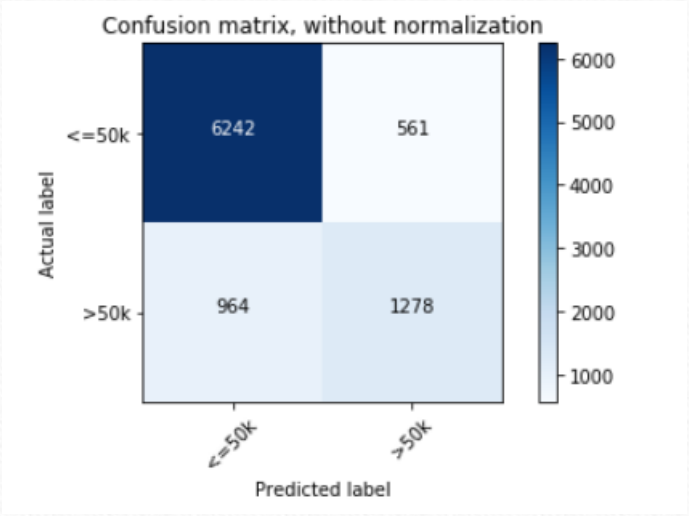
A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

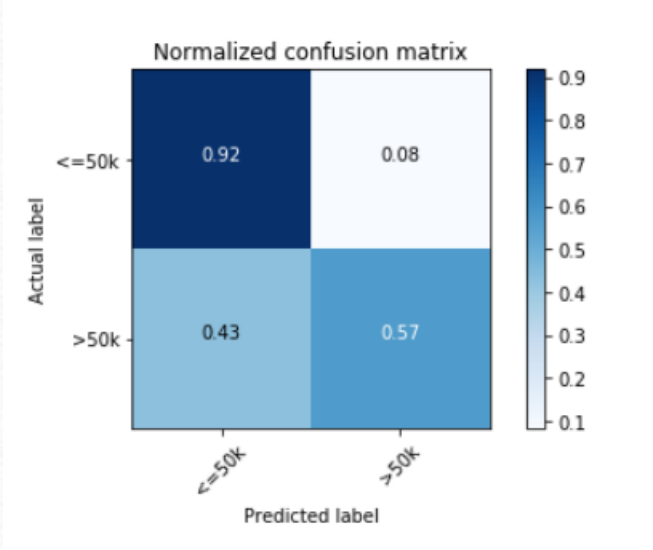
Classification matrix for SVM with default parameter

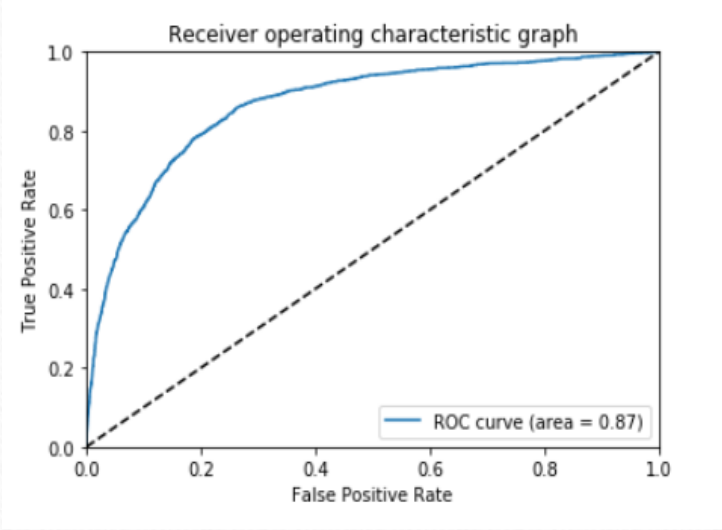


# Classification report, AUC for SVM after hypertuning C and Gamma parameters



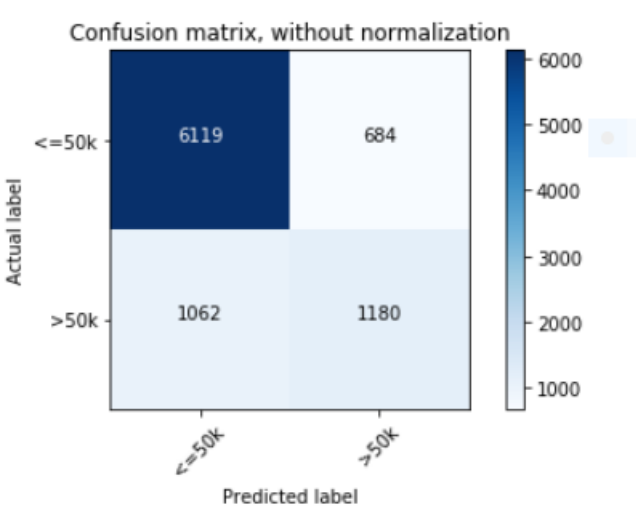



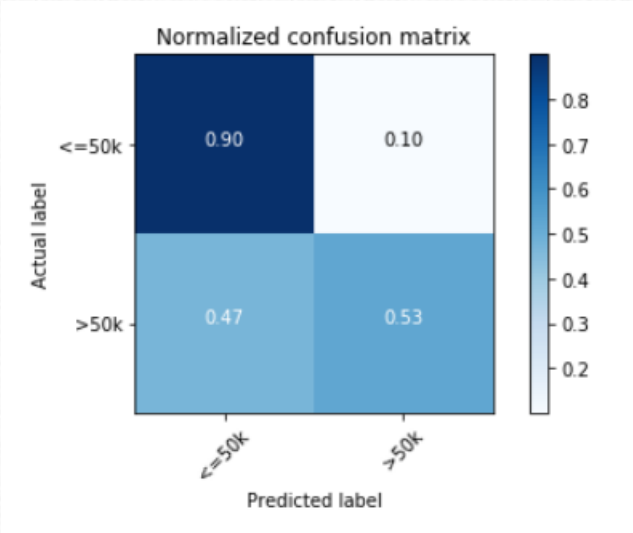



# Random Forest Model

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

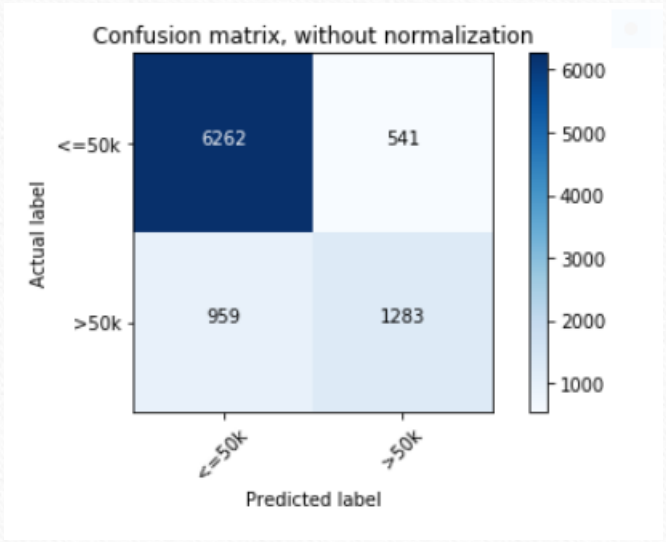
Classification report for Random Forest with default parameters

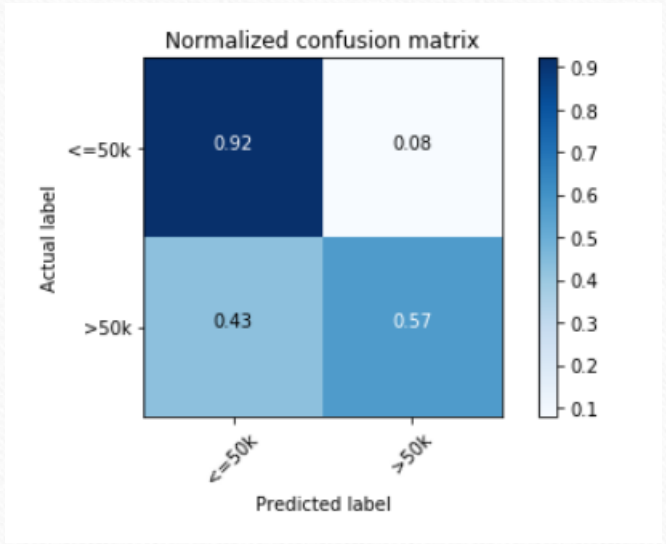


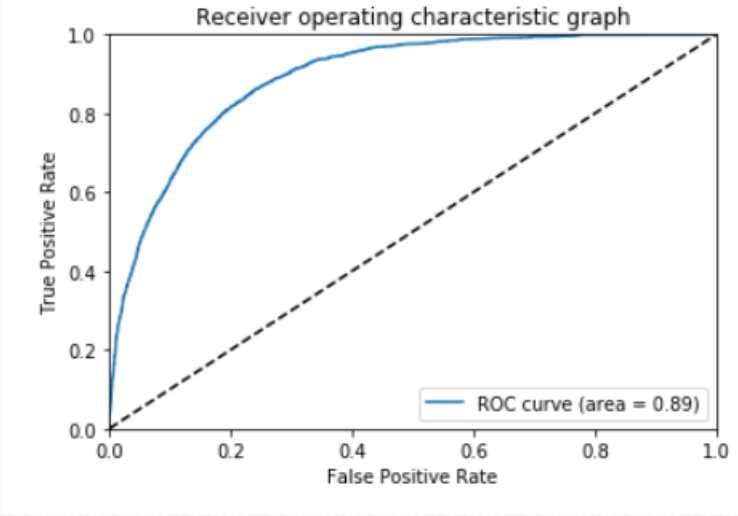



The Random Forest Classifier's parameters tuned to increase the predicting power of the model. Parameters tuned here are max_features, n_estimators, min_sample_leaf

**Classification report and AUC for Random Forest with best params**



## Results of different models

| Model/Metric | Accuracy score | Precision | | Recall | | AUC |
|---|---|---|---|---|---|---|
| | | <=50k | >50k | <=50k | >50k | |
| SVM (default Parameters) | 0.8281 | 0.85 | 0.71 | 0.93 | 0.51 | n/a |
| SVM (gamma = 0.1) | 0.8313 | 0.87 | 0.69 | 0.92 | 0.57 | 0.86 |
| Random Forest (default parameter) | 0.8069 | 0.85 | 0.63 | 0.90 | 0.53 | n/a |
| Random Forest (Best parameters) | 0.8341 | 0.87 | 0.70 | 0.92 | 0.57 | 0.89 |

# Conclusion

- The performances of SVM with best param and Random Forest with best param model are pretty close to each other with accuracy rate of Random Forest model little bit higher, which won't make a big difference in smaller datasets.
- SVM with best param performing at accuracy rate of 83.13%, Random Forest with best params at 83.4%, which means Random Forest with best params model is predicting the <=50k as <=50k classes and >=50k as >=50k classes 0.3% better than SVM with best param model.
- SVM has precision value for Class 1 as 69%, Random forest has slightly improved precision value of 70%. Comparing both models in marketing campaign, less people will be sent irrelevant campaign material with Random forest model.
- SVM model has AUC of 0.86, Random Forest model has 0.89, meaning Random forest model is 3% better than SVM model at separating the classes.

# Further Steps

To improve the model, should try to get more details on the features, Capital loss and Capital gain and build the model including those features. Also, parameters which are not tuned here can be tuned for better performance.